Detección de Eventos Anómalos en Vídeo

Neptalí Menejes Palomino

Orientador: Dr. Guillermo Cámara Chávez

Jurado:
Dr. David Menotti – Universidade Federal do Paraná – Brasil
Dr. Manuel Loaiza – Pontificia Universidade Católica do Rio de Janeiro – Brasil
Dr. Cesar Beltran – Pontificia Universidad Católica del Perú – Perú
Dr. Alex Cuadros – Universidad Católica San Pablo – Perú

Tesis presentada al
Centro de Investigación e Innovación en Ciencia de la Computación (RICS)
como parte de los requisitos para obtener el grado de
Maestro en Ciencia de la Computación.

Universidad Católica San Pablo – UCSP
Marzo de 2017 – Arequipa – Perú

A Dios, por todo lo que me ha dado, a mi madre, mi padre, mis hermanos(as) por su apoyo y amor incondicional y a mis amigos.

Abreviaturas

HMM *Hidden Markov Model*

MDT *Mixtures of Dynamic Textures*

BoW *Bag-of-Words*

HFST *High-Frequency and Spatio-Temporal*

LDA *Latent Dirichlet Allocation*

SFM *Social Force Model*

LBP *Local Binary Patterns*

GMM *Gaussian Mixture Models*

SRC *Sparse Reconstruction Cost*

MHOF *Multi-scale Histogram of Optical Flow*

PSO *Particle Swarm Optimization*

HOP *Histogram of Oriented Pression*

CDI *Crowd Distribution Index*

SIFT *Scale Invariant Feature Transform*

HOOF *Histograms of Oriented Optical Flow*

HOG *Histograms of Oriented Gradients*

HOFG *Histogram of Optical Flow Gradients*

SVM *Support Vector Machines*

OC-SVM *One-Class SVM*

EER *Equal Error Rate*

AUC *Area Under Curve*

ROC *Receiver Operating Characteristic*

TPR *True Positive Rate*

FPR *False Positive Rate*

FNR *False Negative Rate*

UCSD *University of California, San Diego*

UMN *University of Minnesota*

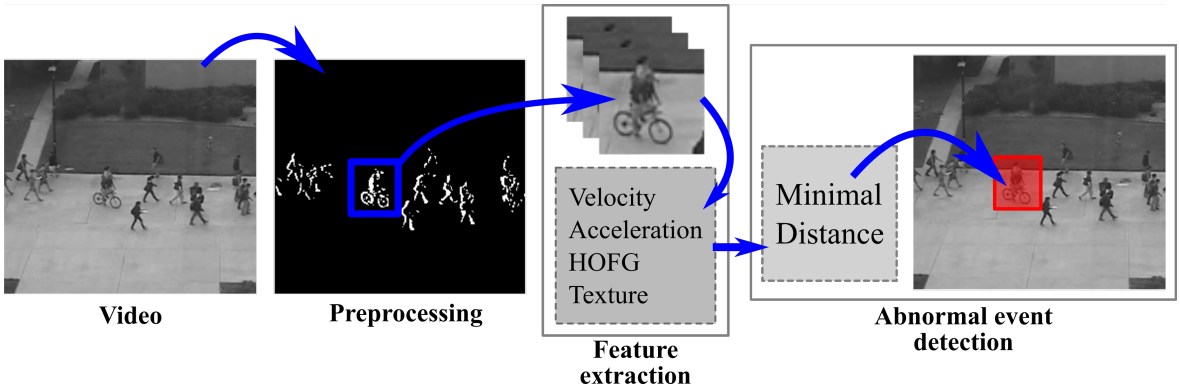
Agradecimientos

En primer lugar deseo agradecer a Dios por haberme guiado a lo largo de estos años de estudio. Agradezco a mis padres y mis hermanos (as) por el apoyo brindado para forjarme como un profesional.

Agradezco de forma muy especial a mi orientador Dr. Guillermo Cámara Chavez por haberme guiado en el desarrollo de esta tesis. Sin su asesoramiento y su retroalimentación no habría sido capaz de terminar esta tesis. También, deseo agradecer al Dr. Alex Cuadros Vargas por sus consejos para seguir y terminar este trabajo de investigación.

Deseo agradecer de manera especial al Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica (CONCYTEC) y al Fondo Nacional de Desarrollo Científico, Tecnológico e Innovación Tecnológica (FONDECYT-CIENCIACTIVA), que mediante Convenio de Gestión UCSP-FONDECYT Nº 011-2013, han permitido la subvención y financiamiento de mis estudios de Maestría en Ciencia de la Computación en la Universidad Católica San Pablo (UCSP).

Abstract



In recent years, detection of anomalous events in video sequences has attracted more attention in the computer vision research community. This has occurred due to the growing need for automated surveillance systems to improve safety in public and private spaces. While progress has been made, there are still some limitations in current research. That is, most of the methods focus on the detection of specific abnormal events, and some are not able to detect more than two types of anomalies.

In this research, a new model for the detection and localization of abnormal events in pedestrian areas is proposed. The goal is to design an algorithm to detect abnormal events in video sequences using motion and appearance information. The motion information is represented through the use of the velocity and acceleration of optical flow and the appearance information is represented by texture and optical flow gradient. To represent these features the use of spatio-temporal patches without overlapping is introduced. Unlike literature methods, proposed model provides a general solution to detect both global and local anomalous events. In addition, the detection stage presents problems of perspective, this is due to the objects near the camera appear to be large, while objects away from the camera appear to be small. To address these problems, classification by region is proposed.

Experimental results on two datasets (UCSD and UMN) and comparison to the state-of-the-art methods validate the performance and robustness of the proposed model. The results of the proposed method on the UCSD Peds2 dataset achieve a EER of 07.2 % and an AUC of 0.977 and in the UMN dataset achieve a 0.998 of AUC in scene 1 and 0.995 of AUC in Scene 3, these results outperform the results of the literature. Meanwhile, results on UCSD Peds1 dataset achieve a EER of 29.2 % and an AUC of 0.792 and in the UMN dataset scene 2 achieves an AUC 0.9486, these results are comparable with results of the methods of the state-of-the-art, this happens because these databases present problems of perspective.

Keywords: Abnormal event detection, video analysis, spatiotemporal feature extraction, video surveillance, computer vision, image processing.

Resumen

En los últimos años, la detección de eventos anómalos en secuencias de vídeo ha atraído una mayor atención en la comunidad de investigación de visión por computador. Esto ha ocurrido debido a la creciente necesidad de utilizar los sistemas de vigilancia automatizados para mejorar la seguridad en los espacios públicos y privados. Si bien se han logrado avances, todavía existen algunas limitaciones en la investigación actual. Es decir, la mayoría de los métodos de la literatura se enfocan en la detección de eventos anómalos específicos, y algunos todavía no son capaces de detectar más de dos tipos de anomalías.

En esta investigación, se propone un nuevo modelo para la detección y localización de eventos anómalos en áreas peatonales. El objetivo es diseñar un algoritmo que permita detectar eventos anómalos mediante el uso de la información de movimiento y la apariencia. La información de movimiento se representa a través del uso de la velocidad y la aceleración del flujo óptico, y la información de apariencia es representado mediante la textura y la gradiente del flujo óptico. Para representar estas características se introduce el uso de parches espacio-temporales sin superposición. A diferencia de los métodos de la literatura, el modelo propuesto proporciona una solución general para detectar eventos anómalos tanto globales como locales. Además, en la etapa de detección se presentan problemas de perspectiva, esto debido a que los objetos cercanos a la cámara parecen ser grandes, mientras que los objetos alejados a la cámara parecen ser pequeños. Para abordar estos problemas, se propone la clasificación por región.

Los resultados experimentales sobre dos bases de datos (UCSD y UMN) y la comparación con los métodos de la literatura validan el rendimiento y la robustez del modelo propuesto. Los resultados del método propuesto sobre la base de datos UCSD Peds2 logra un EER de 07.2 % y un AUC de 0.977 y en la base de datos UMN se logra un 0.998 de AUC en la escena 1 y 0.995 de AUC en la escena 3, estos resultados superan a los resultados de la literatura. Mientras tanto, los resultados sobre las bases de datos UCSD Peds1 logra un EER de 29.2 % y un AUC de 0.792 y en la base de datos UMN escena 2 se logra un 0.948 de AUC, estos resultados son

comparables con los resultados de los métodos de la literatura, esto ocurre debido a que estas bases de datos presentan problemas de perspectiva.

Palabras clave: Detección de eventos anómalos, análisis de vídeo, extracción de características espacio-temporales, videovigilancia, visión por ordenador, procesamiento de imágenes.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

En los últimos años, debido a la creciente necesidad de protección de las personas y propiedades personales, la videovigilancia se ha convertido en una gran preocupación de la vida cotidiana. Una consecuencia de estas necesidades ha llevado a la instalación de cámaras de vigilancia en muchos espacios públicos y privados, tales como: aeropuertos, estaciones de buses, bancos, centros urbanos, o centros comerciales, entre otros. Esto ha ocurrido principalmente debido al aumento de la delincuencia y al miedo que siente la población, por ejemplo según el INEI en el Perú el 31,1 % de la población mayores de 15 años han sido víctimas de un hecho delictivo (INEI, 2016). Si por un lado, una cámara de vigilancia proporciona información visual en tiempo real cubriendo grandes áreas, por otra parte, el número de imágenes adquiridas en un solo día puede ser fácilmente en el orden de miles de millones, lo que complica el almacenamiento de todos los datos e impide su procesamiento.

El análisis automático de los vídeos de vigilancia es un campo importante de la investigación en el área de la visión por computador. Una de las áreas más activas es la comprensión de las actividades por parte del sistema de videovigilancia (Amin et al., 2014). La comprensión de las actividades implica ser capaz de detectar y clasificar el objeto de interés y analizar su comportamiento. Un aspecto fundamental es detectar y reportar situaciones de especial interés, en particular cuando ocurren eventos inesperados. En este caso, un sistema de videovigilancia que puede interpretar la escena y automáticamente reconocer eventos anómalos puede desempeñar un papel vital.

La detección de eventos anómalos es un tema importante de investigación en el sistema de videovigilancia visual. En la literatura existen algunas técnicas propuestas que trabajan en entornos de vigilancia controlados y muestran buenos resultados si están ajustados para una aplicación específica. Sin embargo, estas técnicas todavía no se acercan a un sistema de videovigilancia real que pueda interpretar de forma automática toda la escena y alertar en caso de cualquier situación sospechosa a los operadores o usuarios. Idealmente se espera que sean incluida algún tipo de información semántica con respecto al evento detectado. Además, dependiendo de las aplicaciones, los eventos anómalos se clasifican en dos categorías (Li et al., 2015): la detección de

Evento Anómalo Global (EAG), donde se enfoca en detectar los cambios o eventos basados en la movimiento aparente de toda la escena; y la detección de Evento Anómalo Local (EAL), donde se enfoca en distinguir el EAL que es diferente de sus vecinos espacio-temporales y determinar el lugar donde esta ocurriendo el evento anómalo.

En esta investigación, un nuevo modelo para la detección de eventos anómalos (EAG y EAL) en áreas peatonales es propuesto. El modelo utiliza la información de movimiento y apariencia para representar los eventos anómalos. Los eventos anómalos por lo general suelen tener una velocidad superior (uso de información de movimiento) a los peatones, pero muchas veces estos eventos poseen la misma velocidad que los peatones lo cual hace la detección más difícil. Para abordar este problema se recurre a utilizar la información de apariencia de los objetos anómalos. El modelo propuesto trata de resolver las limitaciones de los enfoques existentes mediante el aprovechamiento de las informaciones de movimiento y apariencia. Una gran parte de este trabajo se dedica al análisis visual del comportamiento humano y la detección del evento anómalo. En este contexto, el modelo propuesto debería alertar situaciones sospechosas, peligrosas para ayudar a mejorar la seguridad pública. Además, en la etapa de detección se presentan problemas de perspectiva, esto debido a que los objetos cercanos a la cámara parecen ser grandes, mientras que los objetos alejados a la cámara parecen ser pequeños. Para abordar estos problemas y obtener mejores resultados, se propone una clasificación por regiones locales.

1.1. Motivación y Contexto

En la actualidad, una gran cantidad de datos de vigilancia se acumulan cada día. Estos son monitoreados por muy pocos observadores en relación a la gran cantidad de cámaras, lo que hace que sea difícil de detectar y responder a todos los eventos anómalos que ocurren en la escena. Además, el análisis y comprensión de los eventos anómalos en una secuencia de vídeo es a la vez un problema científico difícil, y un dominio relativamente nuevo, el cual está atrayendo la atención de muchos investigadores, instituciones y empresas comerciales.

Existen enfoques propuestos en el estado del arte con resultados buenos, pero aún no se acercan a un sistema de detección ideal debido a muchas restricciones en el diseño de software con un costo computacional muy elevado, los cuales limitan el rendimiento del algoritmo. Por lo tanto, un sistema inteligente que realize la detección de eventos anómalos en vídeo es útil para muchos aplicaciones de seguridad.

Analizar correctamente las escenas llenas de gente requiere tener información contextual con respecto a esa escena. Un evento que es anómalo en un contexto, puede considerarse normal en otro. Por ejemplo, en una zona donde hay un carril para las bicicletas, la presencia de una bicicleta se considera normal, mientras que en una zona que es estrictamente peatonal, la existencia de una bicicleta se considera como una anomalía. El contexto de esta investigación está centrado en la detección de eventos

anómalos en áreas peatonales.

1.2. Objetivos

Esta investigación tiene como objetivo diseñar un modelo que permita la detección y localización de eventos anómalos en secuencias de vídeos mediante el uso de la información de movimiento y apariencia.

1.2.1. Objetivos Específicos

- Estudiar los descriptores visuales basados en movimiento y apariencia.
- Diseñar e implementar un algoritmo para la extracción de características de movimiento y apariencia.
- Validar los resultados del método propuesto comparándolos con las anotaciones (ground truth) de las base de datos.

1.3. Contribuciones

Las contribuciones que hacemos en esta investigación se resumen de la siguiente manera:

- Se presenta un modelo que proporciona una solución general para detectar eventos anómalos tanto globales como locales.
- El método propuesto utiliza la combinación de las características de movimiento y apariencia con el objetivo de superar los resultados del estado del arte.
- La fase de entrenamiento solo se requiere de cuadros (*frames*) normales, es decir, de vídeos que contienen eventos normales. Esto evita la necesidad de adquisición de escenas que contengan eventos anómalos, que a menudo son difíciles de obtener en escenas reales.
- Se plantea la clasificación por regiones locales para superar los problemas de perspectiva que presentan algunas secuencias de vídeo.

1.4. Organización de la tesis

El resto de la tesis esta organizado como sigue: En el Capítulo II, una revisión de la literatura de las técnicas de detección de anomalía es proporcionado. En el Capítulo III, se definen algunos conceptos relevantes para el desarrollo de la presente investigación. El método propuesto para la detección de eventos anómalos es explicado a detalle en el Capítulo IV. En el Capítulo V, los resultados experimentales del método propuesto y la comparación con los métodos de la literatura es proporcionado. Finalmente, las conclusiones, las limitaciones y los trabajos futuros se describen en el Capítulo VI.

Capítulo 2

Trabajos Relacionados

En el presente capítulo se presenta una breve revisión de la literatura relevante para el desarrollo de la tesis. En los últimos años, la detección de eventos anómalos ha atraído una gran atención de la investigación en visión por computador. Una variedad de métodos y técnicas han sido propuestas con el objetivo de detectar eventos anómalos. Por lo general, dependiendo de la aplicación específica y el contexto, los métodos de la literatura se pueden categorizar en dos clases (Li et al., 2015): detección de Evento Anómalo Global (EAG) y detección de Evento Anómalo Local (EAL). La primera clase se enfoca en determinar si la escena contiene evento anómalo o no, y la segunda determina la posición donde está ocurriendo el evento. En las siguientes secciones revisamos los trabajos relacionados presentados en cada una de las clases.

2.1. Detección de Evento Anómalo Global

Por lo general, los efectos de auto-organización que ocurre en escenas de multitud de personas resultan en patrones de movimiento regulares. Sin embargo, cuando ocurren los eventos anormales que afectan la seguridad pública, tales como incendios, explosiones, accidentes de transporte, las personas entran en un estado de pánico, esto hace que la dinámica de multitud de personas sea un estado completamente diferente. La detección de EAG pretende distinguir los estados anormales de la multitud de personas de los estados normales. Las metodologías relacionadas existentes por lo general tienden a detectar los cambios o eventos basados en el movimiento aparente estimado de la escena completa. También es importante para un sistema de detección de EAG, no solo detectar de manera eficaz la presencia de evento anómalo en la escena, sino también determinar con precisión el inicio y final de los eventos anómalos.

En la literatura, existen trabajos específicamente para la detección de eventos anómalos globales, a continuación describimos algunos de ellos: Mehran et al. (2009) presentan una nueva manera de formular el comportamiento anormal de una multi-

tud de personas mediante la adopción del modelo de fuerza social, luego utilizan el método *Latent Dirichlet Allocation* (LDA) para detectar la anormalidad. Chen and Huang (2013) proponen un método que considera una región aislada como un vértice y una multitud de personas es representado con un grafo. Para modelar con eficacia las variaciones de la topología, se utilizan características locales, tales como el análisis de sub-grafo basado en valor propio y las deformaciones de triángulo, y las características globales tales como el uso de momentos. Finalmente, ambas características son combinados para detectar si algún evento anómalo está presente en la escena.

Un método para la detección de anomalía global en tiempo real en vídeos con escenas de multitud de personas es propuesto en (Gu et al., 2013), donde utilizan el método SIFT (*Scale Invariant Feature Transform*) para extraer características de movimiento (velocidad), luego el Modelo de Mezcla de Gaussianas (GMM, del inglés *Gaussian Mixture Models*) se adapta sobre el conjunto de datos de entrenamiento. Finalmente, los umbrales para detectar eventos anormales se obtienen automáticamente utilizando las GMM entrenadas.

Recientemente, algunos métodos se han propuesto (Wu et al., 2014; Wang and Snoussi, 2014). Wu et al. (2014) propone un modelo Bayesiano para la detección de comportamiento de escape (huida) de una multitud de personas en vídeos. El movimiento de la multitud de personas son caracterizados utilizando los campos de flujo óptico, y las funciones de densidad de probabilidad de clase condicional del flujo óptico se construyen sobre la base de los atributos del flujo de campo. El comportamiento de escape de multitud de personas se detecta mediante una formulación Bayesiana. Wang and Snoussi (2014) proponen una técnica de extracción de características llamado el Histograma de Flujo Óptico Orientado (HOOF, del inglés *Histograms of Oriented Optical Flow*) para la detección de EAG. Donde primero se extraen características de movimiento utilizando el algoritmo de flujo óptico Horn-Schunk (Horn and Schunck, 1981). Luego, se calcula el HOOF para cada cuadro del vídeo. Finalmente, la clasificación de eventos anormales se realiza utilizando las Máquinas de Vectores de Soporte (SVM, del inglés *Support Vector Machines*).

2.2. Detección de Evento Anómalo Local

Además de la detección de EAG, a menudo necesitamos conocer la ubicación dónde esta ocurriendo el evento anómalo. Los métodos para la detección de EAL se enfocan en determinar la ubicación o el lugar donde está ocurriendo el evento anómalo. Para este objetivo varias técnicas han sido propuestas, las cuales según (Li et al., 2015) se dividen en dos grupos: los enfoques basados en la visión que modelan y predicen los eventos anormales solo utilizando técnicas del área de visión por computador, basados en las características visuales; y los enfoques inspirados en la física (del inglés, *physics-inspired*) que incorporan los modelos físicos para representar la dinámica de la multitud, y lograr la detección de eventos anómalos con métodos de aprendizaje de máquina.

2.2.1. Enfoque basado en Características Visuales

Muchas de las técnicas de aprendizaje de máquina han logrado un gran éxito en las tareas de visión por computador. Las mismas también fueron utilizadas para la detección de eventos anómalos locales. Estos métodos por lo general extraen características visuales y construyen un conjunto de grupos (*clusters*) para representar varios patrones de eventos posibles.

2.2.1.1. Modelo Oculto de Markov:

El Modelo Oculto de Markov (HMM, del inglés *Hidden Markov Model*) es capaz de tomar en cuenta la naturaleza dinámica de las características observadas (Sodemann et al., 2012), es aplicable en la detección de eventos en vídeo así como la detección de eventos anómalos.

Kratz and Nishino (2009) proponen un marco de trabajo basado en HMM para modelar el comportamiento de patrones locales de movimiento espacio-temporales en escenas muy concurridas. La Figura 2.1 muestra un único HMM para cada ubicación espacial de observación. En la fase de entrenamiento, la relación temporal entre los

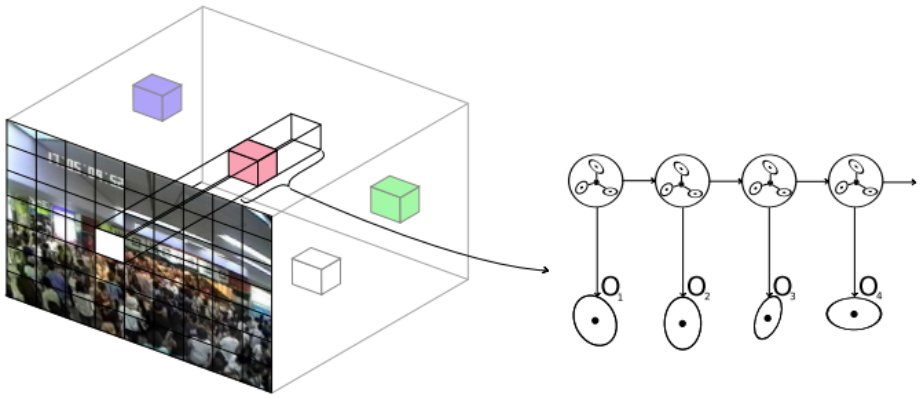


Figura 2.1: La relación temporal entre los patrones locales de movimiento espacio-temporales se codifica con un HMM en cada ubicación espacial. Originalmente mostrado por Kratz and Nishino (2009).

patrones locales de movimiento es capturado a través de un HMM basado en la distribución, y la relación espacial se modela mediante un HMM acoplada. En la fase de prueba, eventos inusuales son identificados como desviaciones estadísticas en las secuencias de vídeo de la misma escena. Los resultados experimentales indican que el modelo propuesto es adecuado para el análisis de escenas muy concurridas. Sin embargo, los autores solo establecen un HMM para cada área local, de modo que el método podría trabajar sólo para tipos limitados de comportamientos normales o escenas concurridas muy específicas. Si se cambia el tipo de comportamiento normal, la tasa de detección de los comportamientos anormales decrecerá, a menos que el modelo sea re-entrenado.

Un esquema similar se ha propuesto en (Wang et al., 2012), en su enfoque, la información de Alta Frecuencia y Espacio-Temporal (HFST, del inglés *High-Frequency and Spatio-Temporal*) es calculado mediante la Transformada de Wavelet para caracterizar las propiedades dinámicas de la región local. Luego, con el objetivo de detectar varios eventos anómalos locales, múltiples HMMs se adaptan, y cada HMM representa un tipo de comportamiento.

2.2.1.2. Modelo de la Textura Dinámica:

La textura dinámica (Chan and Vasconcelos, 2008) es un modelo generativo espacio-temporal para el vídeo, que representa secuencias de vídeo como observaciones de un sistema dinámico lineal y presenta características estacionarias espacio-temporales. Originalmente propuesto para segmentación de movimiento en (Chan and Vasconcelos, 2008), la Mezcla de Texturas Dinámicas (MDT, del inglés *Mixtures of Dynamic Textures*) es un modelo generativo, donde una colección de secuencias de vídeo se modelan como muestras de un conjunto de texturas dinámicas subyacentes. La Figura 2.2 muestra el MDT de un parche del vídeo.

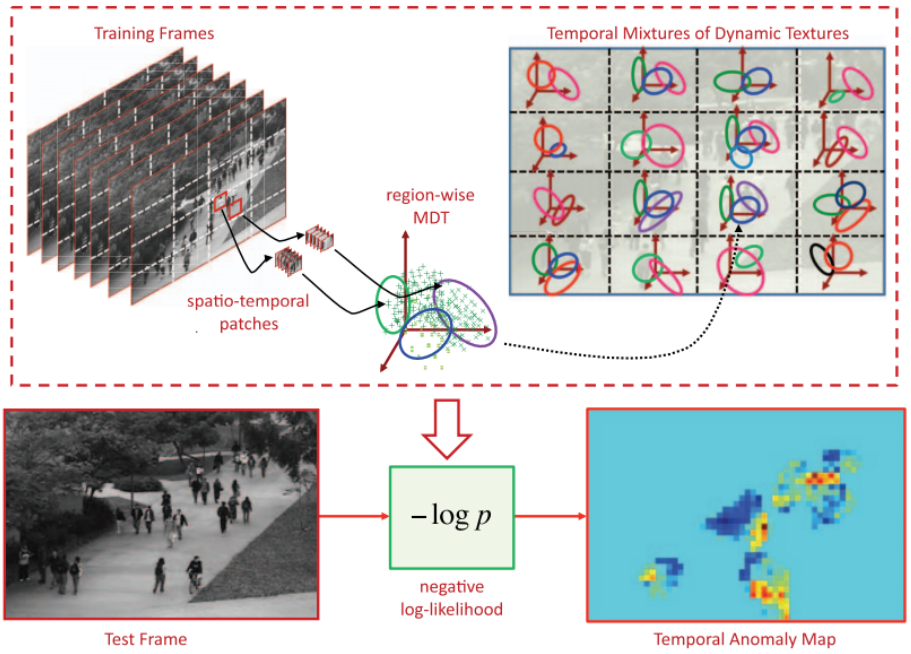


Figura 2.2: MDT para detección de anormalidad temporal. Para cada región de la escena. Originalmente mostrado por Li et al. (2014).

Basado en MDT, un detector conjunto de anomalías temporales y espaciales en escenas muy concurridas es propuesto en (Mahadevan et al., 2010; Li et al., 2014). El detector propuesto se basa en una representación de vídeo que toma en cuenta las características de apariencia y la dinámica de la multitud, utilizando un conjunto de modelos MDT. En la fase de entrenamiento, los patrones normales se aprenden a través

de un modelo MDT por cada subregión de la escena. Un mapa de anomalía temporal multiescala se produce mediante la medición de la probabilidad negativa de cada parche de vídeo bajo el MDT de la región correspondiente. En la fase de prueba, los parches de subregiones de baja probabilidad en el marco del MDT asociado son considerados como anomalías.

2.2.1.3. Modelo de Bolsa de Palabras:

Un enfoque representativo en la detección de anomalías es utilizar volúmenes de vídeo espacio-temporales locales basado en el modelo de Bolsa de Palabras (BoW, del inglés *Bag-of-Words*). Este enfoque generalmente extrae características visuales locales de bajo nivel, tales como movimiento y textura, ya sea mediante la construcción de un modelo de fondo a nivel de píxeles y los patrones de comportamiento, o mediante el empleo de volúmenes de vídeo espacio-temporales.

Roshtkhari and Levine (2013) extienden el modelo BoW para detectar eventos anómalos en los vídeos. El método representa un vídeo como un conjunto compacto de volúmenes espacio-temporales. Las composiciones de los volúmenes espacio-temporales del vídeo se modelan utilizando un marco probabilístico, y los eventos anómalos son considerados como las representaciones de vídeo con muy baja frecuencia de ocurrencia.

LDA se ha adoptado en (Wang et al., 2011) basado en cuboides espacio-temporales de las secuencias de vídeo con un tamaño adaptativo. Para calcular la semejanza entre dos cuboides espacio-temporales con diferentes tamaños, diseñaron un diccionario de palabras visuales (*codebook*). El modelo LDA se utiliza para aprender un número apropiado de tópicos para representar estos escenarios. Una nueva muestra se clasifica como evento anómalo si este no pertenece a estos tópicos.

2.2.1.4. Modelos de Representación Disperso:

Recientemente, Cong et al. (2013) presentó un nuevo algoritmo para la detección de eventos anormales basado en el método de Costo de Reconstrucción Dispersa (SRC, del inglés *Sparse Reconstruction Cost*). Dada una secuencia de imágenes o una colección de parches locales espacio-temporales, se extraen las características mediante la creación de los Histogramas Multi-escala del Flujo Óptico (MHOF del inglés *Multi-scale Histogram of Optical Flow*), donde el histograma se calcula en múltiples escalas del flujo óptico. Luego, se utiliza SRC sobre las muestras de entrenamiento normales para medir la normalidad de las muestras de prueba.

Combinando con la textura dinámica, (Xu et al., 2011) proponen un nuevo enfoque para la detección de eventos inusuales basado en el error de reconstrucción disperso sobre un conjunto base aprendido de las texturas dinámicas, donde este conjunto esta formado solamente con los eventos normales. La textura dinámica es representada por los Patrones Binarios Locales (LBP, del inglés *Local Binary Patterns*) de tres pla-

nos ortogonales. En el proceso de la detección, dado el conjunto base aprendido en el proceso de entrenamiento y la observación de las muestras de entrada, se calculan los coeficientes escasos y se define el error de reconstrucción. Los eventos inusuales se identifican como aquellas texturas dinámicas con elevado error de reconstrucción.

2.2.2. Enfoques Inspirados en la Física

Muchos modelos inspirado en la física se han propuesto para la representación de la multitud, y también han sido utilizados y combinados con técnicas de aprendizaje de maquina para la detección de eventos anómalos. Por ejemplo, el enfoque basado en la serie continua y el enfoque basado en el agente del campo de simulación de la multitud ambos se han adoptado para la detección de eventos anómalos en las escenas muy concurridas.

2.2.2.1. Modelo de Campo de Flujo:

Por lo general, necesitamos entender cómo las multitudes evolucionan con el tiempo y tratar de encontrar algunos patrones regulares. De esta forma es posible conocer inmediatamente dónde y cómo el patrón de movimiento de la multitud de personas esta cambiando.

Wu et al. (2010) propone un método para modelar el flujo de la multitud y la detección de anomalía en las escenas estructuradas y no estructuradas. El flujo de trabajo general comienza con la advección de partículas basado en el flujo óptico, y las trayectorias de las partículas se agrupan para obtener trayectorias representativas del flujo de la multitud. A continuación, la dinámica caótica de todas las trayectorias representativas son extraídas y cuantificadas utilizando invariantes caóticas. Esto se conoce como exponente máximo de Lyapunov y dimensión de correlación en sistema dinámico. El modelo de probabilidad se aprende de estos conjuntos de características caóticas. Finalmente, un criterio de estimación de máxima probabilidad es adoptado para identificar un vídeo de consulta de una escena como normal o anormal.

2.2.2.2. Modelo de Fuerza Social:

El Modelo de Fuerza Social (SFM, del inglés *Social Force Model*) se ha empleado con éxito en campos de investigación como la simulación y el análisis de las multitudes. Mehran et al. (2009) introdujeron un nuevo método para detectar y localizar comportamientos anormales en vídeos con multitud de personas utilizando el modelo SFM. Para este propósito el método propuesto por (Ali and Shah, 2007) es utilizado para calcular el flujo de partículas y sus fuerzas de interacción es estimado utilizando el modelo SFM. Luego la fuerzas de interacción son mapeados en el plano de la imagen para obtener el

flujo de fuerza para cada píxel de cada cuadro. Volúmenes espacio-temporales seleccionados al azar de los flujos de fuerza se utilizan para modelar el comportamiento normal de la multitud. Finalmente, los cuadros se clasifican como normal o anormal mediante el uso del modelo BoW. Las regiones anómalas en el cuadro anormal se localizan utilizando las fuerzas de interacción.

Inspirado por Mehran et al. (2009), algunos métodos basados en SFM para detectar comportamientos anormales de la multitud de personas fue propuesto por (Raghavendra et al., 2011a,b), donde introduce el método Optimización de Enjambre de Partículas (PSO, del inglés *Particle Swarm Optimization*) para optimizar las fuerzas de interacción calculados utilizando SFM. El objetivo principal del método propuesto es desplazar la población de partículas hacia las zonas donde existe mayor movimiento en la imagen. Tales desplazamientos se conducen a través de la función PSO, que tiene como objetivo reducir al mínimo la fuerza de interacción, así como modelar el comportamiento de las multitudes más difundido y típico.

2.2.2.3. Modelo de Energía de Multitud de Personas:

Las multitudes de personas tienen sus propias características. Yang et al. (2012) proponen un modelo de presión local para detectar la anormalidad en las escenas muy concurridas basada en características de la multitud local. Estas características incluyen la densidad local y la velocidad, que son parámetros muy importantes para medir la dinámica de las multitudes. El modelo SFM y los LBP se adoptan para calcular la presión local. La Histograma de la Presión Orientada (HOP, del inglés *Histogram of Oriented Pression*) se utiliza para extraer la propiedad estadística de la magnitud y dirección de la presión. Finalmente, se utiliza el clasificador SVM para detectar la anomalía.

Xiong et al. (2011) proponen un nuevo método para detectar dos actividades anormales típicas: concentración y escape de peatones. El método esta basado en la energía potencial y la energía cinética. Un término llamado Índice de Distribución de Multitud (CDI, del inglés *Crowd Distribution Index*) es definido para representar la dispersión, que más tarde puede determinar la energía cinética. Finalmente las actividades anormales se detectan a través de análisis de umbral.

2.3. Consideraciones finales

Según la literatura los métodos y técnicas planteados son evaluados en diferentes contextos y bases de datos. Esto hace que la comparación entre ellos sea aún más difícil. También podemos observar según la literatura que los métodos basados en apariencia, específicamente las texturas dinámicas, obtienen buenos resultados en cuanto a precisión. Sin embargo, estos métodos generan un costo computacional muy elevado. Mientras tanto los métodos basados en flujo óptico para la representación de la multi-

tud son más convenientes para la detección de eventos anómalos en vídeos, obteniendo buenos resultados con menos costo computacional.

Capítulo 3

Fundamentos Teóricos

El procesamiento de vídeo por lo general se divide en dos etapas: la etapa de extracción de características, en esta etapa se realiza principalmente la descripción de las características visuales tales como la apariencia, el movimiento, etc., mediante el uso de los diferentes descriptores visuales; y la etapa de detección de eventos en secuencias de vídeo, mediante el uso de clasificadores como técnicas de aprendizaje de máquina. En las secciones siguientes explicaremos algunas técnicas utilizadas en cada una de las etapas mencionadas. En la Sección 3.1 explicaremos sobre los conceptos y aplicaciones de los descriptores visuales y en la Sección 3.2 explicamos el uso de clasificadores para la detección de eventos en secuencias de vídeo.

3.1. Descriptores Visuales

Los descriptores visuales describen las características visuales de los contenidos dispuestos en imágenes o en vídeos. Describen características elementales tales como la forma, el color, la textura o el movimiento, entre otros. Estas características por los general se clasifican en dos grupos: las características basadas en apariencia y las características basados en movimiento. En la literatura existen técnicas o algoritmos que producen tales descripciones, por ejemplo para la descripción de las características de movimiento se utilizan los algoritmos del flujo óptico, y para la detección de objetos basados en las características de apariencia, son utilizados los Histogramas de Gradientes Orientados (HOG, del inglés *Histograms of Oriented Gradients*). A continuación definimos algunos descriptores relevantes para el desarrollo de la presente tesis.

3.1.1. Flujo Óptico

EL flujo óptico es la distribución de las velocidades aparentes de movimiento de los patrones de brillo en una imagen (Horn and Schunck, 1981). Se refiere al movimiento aparente de los píxeles de un cuadro a otro dentro de una secuencia de vídeo, y puede ser considerado como la distancia de movimiento de un píxel en dos cuadros consecutivos. En la Figura 3.1 se muestra el flujo óptico de una secuencia de vídeos en dos escenarios distintos.

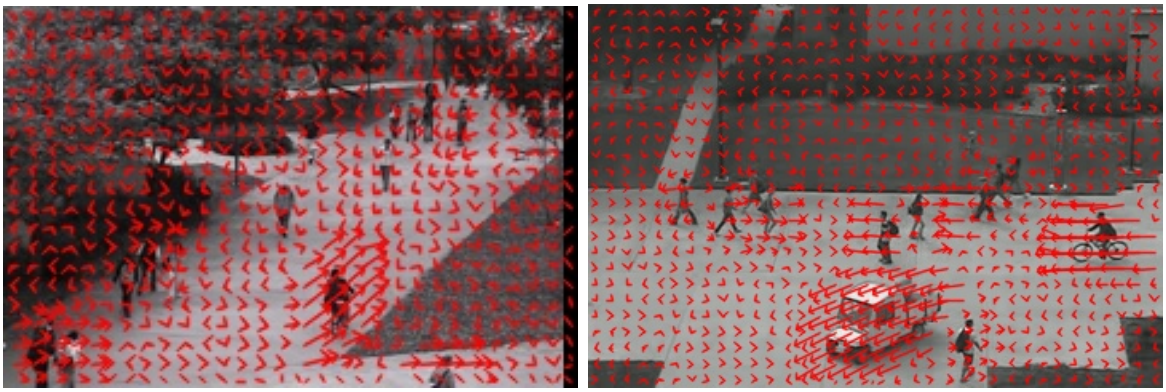


Figura 3.1: Flujo óptico de una secuencia de vídeos en dos escenarios distintos.

El enfoque del flujo óptico generalmente se utiliza para estimar los movimientos de objetos en primer plano (*foreground*) de la escena. El flujo óptico posee varias ventajas en la estimación de los movimientos de objetos en primer plano. En primer lugar, sólo se requieren dos cuadros consecutivos para estimar los movimientos de los píxeles en la escena, lo que permite una respuesta más rápida a los cambios en la velocidad de movimiento y dirección. En segundo lugar, el método de flujo óptico es capaz de descubrir los movimientos parciales o locales dentro de los objetos, por ejemplo los movimientos de las manos y las piernas; esta propiedad hace que el flujo óptico sea un método más factible en la detección de interacciones entre objetos. Finalmente, el flujo óptico es un método relativamente eficaz, por lo que los movimientos de objetos se pueden extraer en tiempo real.

3.1.1.1. Estimación de Flujo Óptico

El enfoque de flujo óptico se utiliza para estimar el movimiento aparente de los píxeles entre dos cuadros consecutivos en un tiempo t y $t+\Delta t$ en cada posición espacial (x, y) del cuadro. Esta forma de estimar el movimiento se denomina como un método diferencial ya que se basan en aproximaciones locales de la serie de Taylor de la señal de la imagen. Es decir, utilizan las derivadas parciales con respecto a las coordenadas espaciales y temporales.

Sea la intensidad de un píxel en la coordenada espacial (x, y) y en el tiempo t,

lo cual denotamos como $I(x, y, t)$. La restricción de brillo se basa en la suposición de que el brillo del píxel I se mantiene constante cuando se mueve un desplazamiento de $(\Delta x, \Delta y, \Delta t)$.

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{3.1}$$

$$I(x, y, t) = I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t + \epsilon \tag{3.2}$$

donde ϵ representa los términos de segundo orden y de orden superior. Para un movimiento relativamente pequeño $(\Delta x, \Delta y, \Delta t)$, el valor de ϵ es pequeño y se puede representar como,

$$\frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t = 0 \tag{3.3}$$

ó

$$\frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t}\frac{\Delta t}{\Delta t} = 0. \tag{3.4}$$

Por razones de brevedad el lado izquierdo de la ecuación anterior se representa de la siguiente forma:

$$I_x u + I_y v + I_t = 0, \tag{3.5}$$

donde $u = \frac{\Delta x}{\Delta t}$ y $v = \frac{\Delta y}{\Delta t}$ son las variables desconocidas del flujo óptico, y $I_x = \frac{\partial I}{\partial x}$, $I_y = \frac{\partial I}{\partial y}$ y $I_t = \frac{\partial I}{\partial t}$ representan las derivadas horizontal, vertical y temporal de la imagen respectivamente. Estos gradientes son generalmente estimados utilizando la diferencia entre los píxeles adyacentes.

Como la Ecuación 3.5 contiene dos variables desconocidas que no se puede resolver como tal, son necesarios algunas restricciones adicionales. Este problema se conoce como el problema de la abertura ya que la información de movimiento local es ambigua cuando se ve a través de una pequeña ventana (o abertura). Los algoritmos de flujo óptico introducen restricciones adicionales para resolver este problema.

3.1.1.2. Estimación del Flujo Óptico utilizando el método de Horn-Schunck

Horn and Schunck (1981) proponen un algoritmo introduciendo una restricción global de suavidad para calcular el flujo óptico basado en la intuición de que los píxeles vecinos tienen una velocidad similar. El método Horn-Schunck (HS) combina un término datos con un término espacial. El término datos asume la constancia de alguna propiedad de la imagen, y el término espacial modela la variación del flujo esperado a través de la imagen. Para secuencias de imágenes de dos dimensiones en escala de grises, el flujo óptico es formulado como una energía global funcional:

$$E = \int\int \left[(I_x u + I_y v + I_t)^2 + \alpha\left(\|\nabla u\|^2 + \|\nabla v\|^2\right)\right] dxdy, \tag{3.6}$$

donde I_x, I_y y I_t son las derivadas de los valores de intensidad de la imagen a lo largo de la dirección horizontal x, dirección vertical y y la dimensión de tiempo t, respectivamente, u y v son los componentes horizontal y vertical del flujo óptico y α es el parámetro que representa el peso del término de regularización. Las ecuaciones

de Lagrange son utilizadas para minimizar la función E, obteniéndose las ecuaciones siguientes:

$$\begin{cases} I_x\left(I_x u + I_y v + I_t\right) - \alpha^2 \Delta u = 0 \\ I_y\left(I_x u + I_y v + I_t\right) - \alpha^2 \Delta v = 0, \end{cases} \tag{3.7}$$

sujeto a

$$\begin{cases} \Delta u\left(x, y\right) = \overline{u}\left(x, y\right) - u\left(x, y\right) \\ \Delta v\left(x, y\right) = \overline{v}\left(x, y\right) - v\left(x, y\right), \end{cases} \tag{3.8}$$

donde $\overline{u}$ y $\overline{v}$ son promedios ponderados de u y v respectivamente, calculados en una vecindad alrededor de la posición del píxel. Como la solución depende de los valores vecinos, cuando se actualizan los píxeles vecinos, la solución necesita ser iterada. El flujo óptico es calculado en un esquema iterativo como se muestra a continuación:

$$\begin{cases} u^{k+1} = \overline{u}^k - \frac{I_x\left(I_x \overline{u}^k + I_y \overline{v}^k + I_t\right)}{\alpha^2 + I_x^2 + I_y^2} \\ v^{k+1} = \overline{v}^k - \frac{I_y\left(I_x \overline{u}^k + I_y \overline{v}^k + I_t\right)}{\alpha^2 + I_x^2 + I_y^2}, \end{cases} \tag{3.9}$$

donde k representa el numero de iteraciones del algoritmo. El número de iteraciones depende de la precisión requerida y la calidad de la estimación inicial. Para las secuencias de vídeo, la estimación inicial es típicamente el campo de flujo óptico a partir del cuadro anterior.

3.1.1.3. Estimación del Flujo Óptico utilizando el método de Lucas-Kanade

El método Lucas-Kanade (LK) (Bouguet, 2001) es uno de los métodos más populares para la estimación de movimiento. El flujo de los píxeles en la imagen se calcula mediante la comparación de las derivadas parciales de las imágenes. En contraste con el método Horn-Schunck que se considera como el algoritmo basada en restricción global, el método LK se basa en una restricción local y el flujo se estima dentro de un parche (ventana) de la imagen utilizando una estimación de mínimos cuadrados. El método LK supera el método de Horn-Schunck cuando hay muchos movimientos locales en lugar de los flujos globales.

El método de LK asume que el desplazamiento de los píxeles de la imagen entre dos cuadros consecutivos es pequeño y aproximadamente constante dentro de una pequeña ventana centrado en (x, y). Considerando que la dimensión de la ventana sea $w \times w$ con $w > 1$, la ecuación del flujo óptico se puede asumir para mantener todos los píxeles dentro de una ventana centrada en (x, y). Es decir, el vector de flujo local de la imagen (u, v) debe satisfacer lo siguiente:

$$\begin{aligned} I_{x1} u + I_{y1} v &= -I_{t1} \\ I_{x2} u + I_{y2} v &= -I_{t2} \\ &\vdots \\ I_{xw^2} u + I_{yw^2} v &= -I_{tw^2} \end{aligned} \tag{3.10}$$

donde $(I_{x1}, I_{x2}, ..., I_{xw^2})$, $(I_{y1}, I_{y2}, ..., I_{yw^2})$ y $(I_{t1}, I_{t2}, ..., I_{tw^2})$ son las derivadas parciales dentro de la ventana centrada en (x, y) y el tiempo t.

En el sistema de ecuaciones anterior tenemos más de dos ecuaciones para las dos incógnitas y por lo tanto el sistema está sobre-determinado. Por otra parte, estas ecuaciones pueden ser representado por un sistema de ecuación lineal $A\overline{v} = b$, donde:

$$A = \begin{bmatrix} I_{x1} & I_{y1} \\ I_{x2} & I_{y2} \\ \vdots & \vdots \\ I_{xw^2} & I_{yw^2} \end{bmatrix}, \quad \overline{v} = \begin{bmatrix} u \\ v \end{bmatrix}, \quad y \quad b = \begin{bmatrix} -I_{t1} \\ -I_{t2} \\ \vdots \\ -I_{tw^2} \end{bmatrix}$$

El método LK calcula el valor del flujo óptico $\overline{v}$ mediante el método de mínimos cuadrados. Es decir, se resuelve un sistema de 2×2.

$$A^T A\overline{v} = A^T b \tag{3.11}$$

La ecuación anterior se puede escribir de otra manera, entonces se tiene:

$$\overline{v} = \left(A^T A\right)^{-1} A^T b \tag{3.12}$$

donde A^T es la transpuesta de la matriz A. El flujo óptico $\overline{v}$ se obtiene calculando el siguiente sistema:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{w^2} I_{xk}^2 & \sum_{k=1}^{w^2} I_{xk} I_{yk} \\ \sum_{k=1}^{w^2} I_{xk} I_{yk} & \sum_{k=1}^{w^2} I_{yk}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{k=1}^{w^2} I_{xk} I_{tk} \\ -\sum_{k=1}^{w^2} I_{yk} I_{tk} \end{bmatrix} \tag{3.13}$$

donde la matriz central en la ecuación es una matriz inversa y la matriz $A^T A$ es conocido como la matriz de momento de segundo orden de la imagen en la posición (x, y).

La solución de mínimos cuadrados en la Ecuación 3.13, otorga la misma importancia a todos los píxeles de la ventana $w \times w$. En la practica por lo general es mejor dar mayor peso a los píxeles que están más cerca del píxel central (x, y). Por ello, se utiliza la versión ponderada de la ecuación de mínimos cuadrados.

$$A^T W A\overline{v} = A^T W b \tag{3.14}$$

ó

$$\overline{v} = \left(A^T W A\right)^{-1} A^T W b \tag{3.15}$$

donde W es una matriz diagonal $w \times w$ que contiene los pesos $W_{kk} = w_k$ que se asignará en la ecuación del píxel I_k. Por lo tanto, el flujo óptico $\overline{v}$ se calcula a partir del siguiente sistema:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \sum_{k=1}^{w^2} w_k I_{xk}^2 & \sum_{k=1}^{w^2} w_k I_{xk} I_{yk} \\ \sum_{k=1}^{w^2} w_k I_{xk} I_{yk} & \sum_{k=1}^{w^2} w_k I_{yk}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum_{k=1}^{w^2} w_k I_{xk} I_{tk} \\ -\sum_{k=1}^{w^2} w_k I_{yk} I_{tk} \end{bmatrix} \tag{3.16}$$

donde el peso w_k generalmente se establece utilizando una función de distancia Gaussiana entre I_k y (x, y).

El método de Lucas-Kanade tradicional no puede detectar el movimiento de alta velocidad; ya que supone que el movimiento es lo suficientemente pequeño en relación al tamaño de la ventana. Para abordar este problema de la detección de movimientos largos o de alta velocidad, el algoritmo original de LK es modificado por (Bouguet, 2001), donde el algoritmo de estimación de flujo se calcula en distintos niveles de escala, conocido como el método de Lucas-Kanade piramidal. La diferencia entre el método LK tradicional y piramidal es ilustrado en la Figura 3.2. Se puede observar que el enfoque piramidal es más preciso en la detección de movimientos más largos.

(a) (b)

Figura 3.2: Flujo óptico estimado por el método de Lucas-Kanade tradicional se muestra en (a) y el flujo óptico estimado por el enfoque piramidal se muestra en (b).

3.1.2. Histograma de Flujo Óptico Orientado

El Histograma de Flujo Óptico Orientado (HOOF del inglés, *Histograms of Oriented Optical Flow*), es un descriptor de características de movimiento presentado inicialmente para la detección de acciones humanas en (Chaudhry et al., 2009). Las características HOOF por lo general son utilizados en la detección de eventos anormales globales. En la Figura 3.3 se muestra un esquema general de extracción de características HOOF.

3.1.2.1. Implementación del Algoritmo

La extracción de HOOF proporciona un histograma $h_{t,b} = [h_{t,1}, h_{t,2}, \cdots, h_{t,B}]$ en cada instante de tiempo t, para cada bloque b en el cuadro, en el cual cada vector de flujo es establecido en un *bin* del histograma de acuerdo a su ángulo primario que es calculado con respecto al eje horizontal y se pondera de acuerdo a su magnitud. Por lo tanto, todos los vectores de flujo óptico $v = [x, y]^T$, con orientación $\theta = \tan^{-1}\left(\frac{y}{x}\right)$ y en

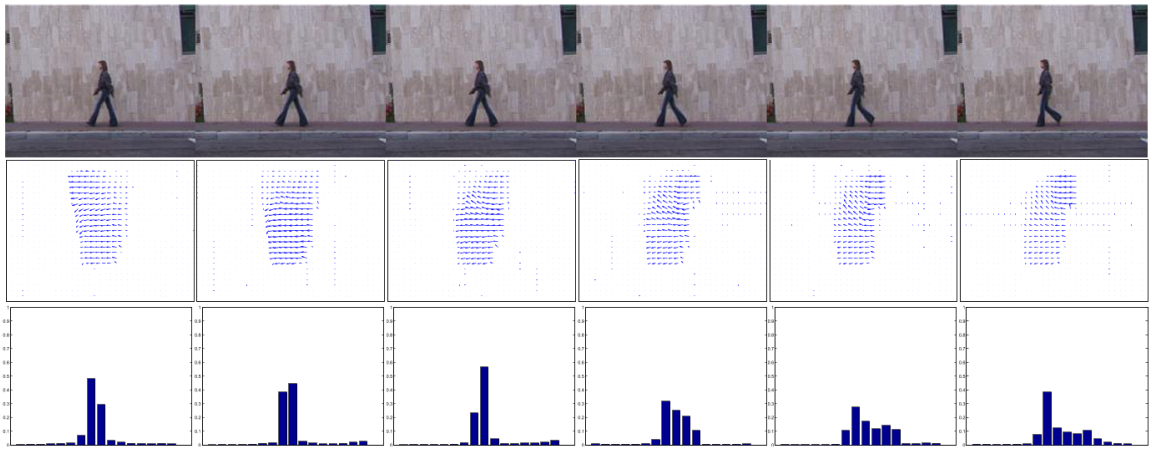


Figura 3.3: El flujo óptico y las características HOOF de una secuencia de imagenes, originalmente mostrado en Chaudhry et al. (2009).

el rango de:

$$-\frac{\pi}{2}+\pi\frac{b-1}{B} \leq \theta < -\frac{\pi}{2}+\pi\frac{b}{B} \tag{3.17}$$

contribuyen con su magnitud $m = \sqrt{x^2+y^2}$ al i-ésimo *bin* del histograma, donde $1 \leq \theta \leq B$, para un total de B *bins*. La Figura 3.4 muestra el procedimiento de la generación del histograma. Finalmente, el histograma es normalizado para hacer que la representación del histograma sea invariante a escala. Dado que la contribución de cada vector de flujo óptico a su *bin* correspondiente es proporcional a su magnitud, el calculo del flujo óptico con presencia de ruidos pequeños tienen poco efecto en el histograma observado.

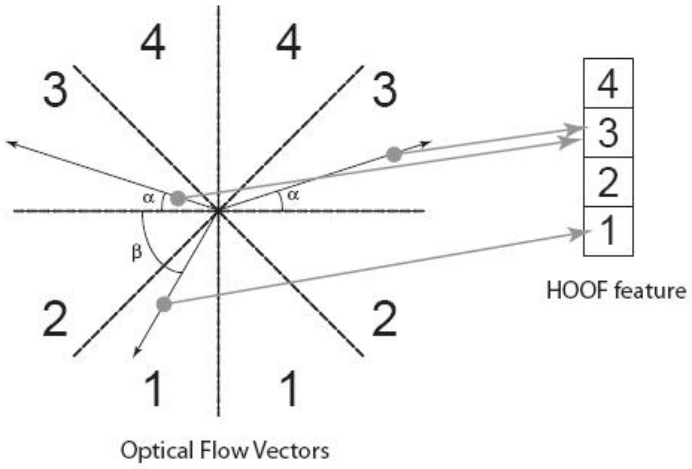


Figura 3.4: Proceso de generación del histograma HOOF, con cuatro *bins*, originalmente mostrado en Chaudhry et al. (2009).

3.1.3. Histograma de Gradientes Orientados

El Histograma de Gradientes Orientados (HOG del inglés *Histograms of Oriented Gradients*) es un descriptor de características utilizado en el ámbito de visión por computador y procesamiento de imágenes con el propósito de detección de objetos en imágenes (Dalal and Triggs, 2005). La idea principal del algoritmo es que la apariencia y la forma local de un objeto en una imagen pueden ser descritas por la distribución de la intensidad de gradiente o la dirección de los contornos. Para esto la imagen es dividida en regiones adyacentes de dimensión pequeña llamadas celdas, y para cada celda dentro de la imagen, se genera un histograma de gradientes orientados. Para obtener mejores resultados, los histogramas locales pueden ser normalizados mediante el cálculo de una medida de la intensidad a través de una región más grande de la imagen, llamado bloque, y luego utilizar este valor para normalizar todas las celdas dentro del bloque. Esta normalización resulta tener una mejor invarianza a los cambios de iluminación y sombras. El descriptor HOG posee algunas ventajas frente a otros descriptores tales como a invariancia a las transformaciones geométricas y fotométricas.

En el trabajo de (Dalal and Triggs, 2005), se enfocaron en la detección de personas en imágenes estáticas, aunque desde entonces expandieron sus pruebas para incluir la detección de peatones en vídeo, así como a una variedad de animales y vehículos comunes en imágenes estáticas.

3.1.3.1. Implementación del Algoritmo

3.1.3.2. Cálculo de Gradiente

La primera etapa del cálculo del descriptor HOG consiste en determinar los gradientes de la imagen. El método más común es la aplicación del filtro de una dimensión (1-D), centrada en una o ambas direcciones horizontal y vertical. Específicamente, este método requiere filtrar los datos de color o intensidad de la imagen con los siguientes máscaras de filtro: $[-1, 0, 1]$ y $[-1, 0, 1]^T$. En el caso de imágenes a color, el gradiente se calcula por separado para cada componente RGB, y para cada píxel se asigna la gradiente con mayor magnitud. La Figura 3.5 muestra la gradiente en las direcciones horizontal y vertical de una imagen en escala de grises.

Otros tipos de máscaras más complejas fueron probadas, tales como el filtro de Sobel 3×3, o máscaras diagonales. Dalal and Triggs (2005) también trataron de aplicar un filtro Gaussiano antes de aplicar la máscara, pero estas operaciones tienen el rendimiento significativamente menor en comparación con la aplicación de la máscara derivada simple y sin filtrado.

(a) (b)

Figura 3.5: Gradiente de una imagen, en ambas direcciones (a) vertical y (b) horizontal

3.1.3.3. Construcción de los Histogramas

La segunda etapa es construir los histogramas de gradientes orientados, siendo generado un histograma para cada celda. Cada píxel de la celda emite un voto ponderado para un *bin* del histograma, dependiendo de la orientación del gradiente calculado en este píxel. Los *bins* del histograma son distribuidos de manera uniforme de 0 a 180° o de 0 a 360° dependiendo si la orientación del gradiente es considerado con o sin signo. Dalal and Triggs (2005) consiguieron mejores resultados con un histograma de 9 *bins* en sus experimentos de detección de personas. Para el voto ponderado también se considera, la magnitud del gradiente de los píxeles, o alguna función de la magnitud. En las pruebas, la magnitud del gradiente generalmente produce los mejores resultados. Otras opciones para el voto ponderado se podrían incluir la raíz cuadrado de la magnitud del gradiente, o alguna versión recortada de la magnitud.

3.1.3.4. Formación de los Bloques

Un paso importante es la estandarización de los descriptores para evitar disparidades debido a las variaciones de iluminación y contraste. Los valores altos del gradiente deben ser localmente normalizado, lo que requiere la agrupación de las celdas adyacentes en bloques conectadas espacialmente. El descriptor HOG es un vector concatenado de los histogramas normalizados de las celdas componentes de todos los bloques de la imagen. Estos bloques normalmente se superponen, lo que significa que cada celda contribuye con su histograma más de una vez al descriptor final. En la Figura 3.6 se muestra el proceso de formación de los bloques.

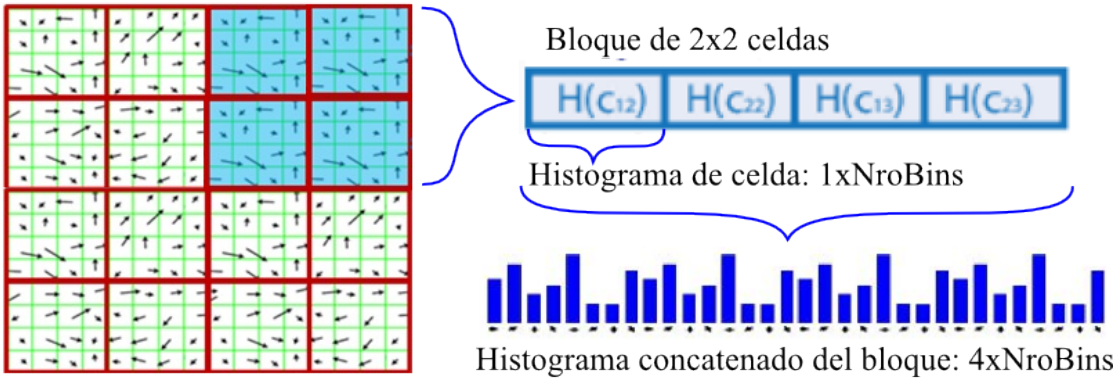


Figura 3.6: Proceso de formación de los bloques

3.1.3.5. Normalización de los Bloques

Se ha evaluado cuatro diferentes métodos de normalización en (Dalal and Triggs, 2005). Sea v un vector sin normalizar que contiene todos los histogramas de un bloque designado, $\|v\|_k$ sea su k-norma para $k = 1, 2$ y e sea un valor constante pequeño. Luego el factor de normalización puede ser uno de los siguientes:

$$L2 - norma : f = \frac{v}{\sqrt{\|v\|_2^2 + e^2}}$$

$$L1 - norma : f = \frac{v}{(\|v\|_1 + e)}$$

$$L1 - sqrt : f = \sqrt{\frac{v}{(\|v\|_1 + e)}}$$

Además, una cuarta esquema $L2 - hys$, puede ser calculado a partir de la $L2 - norma$, este esquema limita los valores máximos de v a 0,2 u a algún otro valor.

3.1.3.6. Clasificación de Descriptores

El paso final en reconocimiento de objetos utilizando descriptores HOG es alimentar los descriptores en algún sistema de reconocimiento basado en el aprendizaje supervisado. Este paso no es parte de la definición del propio descriptor HOG y diversos tipos de clasificadores pueden ser utilizados. Dalal and Triggs (2005) eligen voluntariamente el clasificador SVM, con kernel lineal, para medir fundamentalmente la contribución del descriptor HOG.

3.2. Clasificación

Para la detección de eventos en vídeo, una etapa muy importante es la fase de clasificación. En esta etapa por lo general se recurre al uso de algoritmos de aprendizaje

de maquina. En las siguientes secciones definiremos algunos algoritmos de clasificación más utilizados en la literatura.

3.2.1. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (SVM del ingles, *Support Vector Machines*) es un método basado en la teoría de aprendizaje estadístico y minimización de riesgos para la clasificación y la regresión, inicialmente propuesto por Vapnik and Lerner (1963). Posteriormente, SVM ha sido extendido con la introducción de kerneles no lineales en (Boser et al., 1992; Cristianini and Shawe-Taylor, 2000). La teoría detrás de SVM es brevemente presentado a continuación.

SVM es un sistema de aprendizaje basado en la teoría del aprendizaje estadístico. Supongamos el caso de los clasificadores hiperplano de dos clases en el espacio H:

$$\langle w, x\rangle + b = 0 \qquad w, x \in H \qquad b \in \mathbb{R}$$

correspondiente a la función de decisión:

$$f(x) = sgn\left(\langle w, x\rangle + b\right)$$

La teoría del aprendizaje estadístico indica que el clasificador óptimo se puede encontrar mediante la maximización del margen (Cristianini and Shawe-Taylor, 2000). Esto se puede expresar como un problema de minimización:

$$\min_{w} = \frac{1}{2}\|w\|^2$$

Sujeto a,

$$y_i\left(\langle w, x\rangle + b\right) \geq 1, \qquad i = 1, ..., m$$

donde m es el numero de datos de entrenamiento y $y_i \in \{-1, +1\}$ es la etiqueta de la muestra. SVM construye un hiperplano o conjunto de hiperplanos en un espacio de alta dimension que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá un clasificación correcta, como se muestra en la Figura 3.7.

Supongamos que un conjunto de datos de entrenamiento esta conformado por los muestras de una sola clase (clase positiva), en este caso existe un problema de clasificación de una sola clase (OC-SVM del inglés, *One-Class SVM*). En la siguiente Sección 3.2.2 se explica el concepto de máquinas de vectores de una sola clase para abordar este tipo de problemas.

3.2.2. Máquinas de Vectores de Soporte de Una Clase

En los problemas de detección de eventos anormales, generalmente sólo se dispone de datos o patrones normales conocido también como clase positiva, y los valores

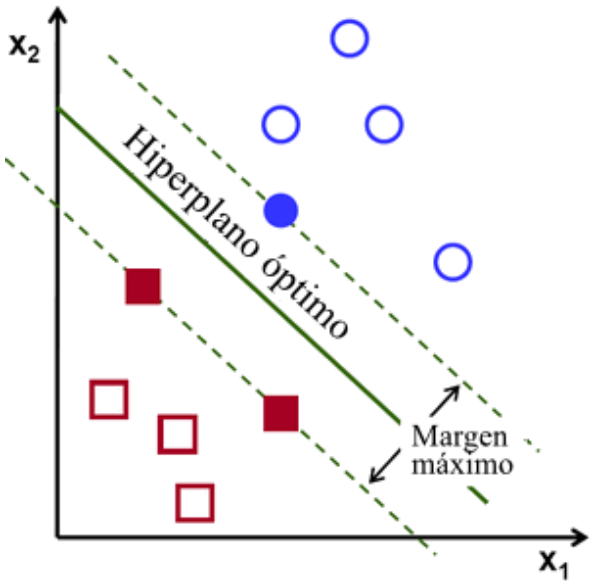


Figura 3.7: Construcción de un hiperplano óptimo con sus margenes maximizados de un conjunto de datos de entrenamiento de dos clases

atípicos o anormales se presentan rara vez. El clasificador *One-Class SVM* es propuesto para abordar este tipo problemas donde sólo las muestras positivas con pocos valores atípicos están disponibles.

One-Class SVM tiene como objetivo determinar una región adecuada en el espacio de datos de entrada X, que incluye la mayoría de las muestras extraídas de una distribución de probabilidad desconocida P. La hiperesfera OC-SVM identifica los valores atípicos mediante el ajuste de una hiperesfera con un radio mínimo. En la Figura 3.8 se muestra la geometría de la hiperesfera *One-Class SVM*.

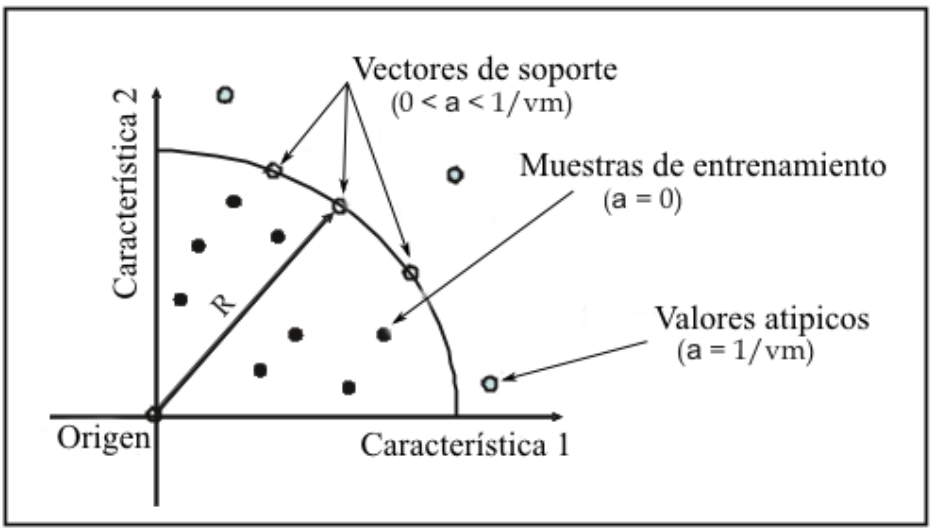


Figura 3.8: Geometría de la hiperesfera *One-Class SVM*.

El hiperplano *One-Class SVM* es una extensión de la versión original del clasificador SVM para abordar los problemas de una sola clase (Schölkopf and Smola, 2002), donde se identifican los valores atípicos mediante el ajuste de un hiperplano desde el origen. La Figura 3.9 muestra un hiperplano *One-Class SVM*. El hiperplano *One-Class SVM* es formulado como el siguiente problema de minimización restringida:

$$\underset{w,\xi,\rho}{min}\frac{1}{2}\|w\|^2 + \frac{1}{vn}\sum_i \xi_i - \rho. \tag{3.18}$$

Sujeto a:

$$\langle w.\Phi(x_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0 \tag{3.19}$$

donde $x_i \in X$, $i \in [1, ..., n]$ son los n muestras de entrenamiento en el espacio de datos de entrada X, y ξ_i es la variable de holgura (*slack variable*) para penalizar a los valores atípicos. El hiperparámetro $v \in (0, 1]$ es el peso para la variable de holgura ξ_i, que ajusta el número de valores atípicos aceptables y permite el análisis de datos con ruido, $\|.\|$ denota la norma Euclidiana de un vector. El hiperplano de decisión está dada por la siguiente ecuación:

$$\langle w, \Phi(x_i)\rangle - \rho = 0. \tag{3.20}$$

La función no lineal $\Phi : X \rightarrow H$, mapea la muestra de entrada x_i desde el espacio de datos de entrada X, al espacio de características H, que nos permite resolver un problema de clasificación no lineal mediante el diseño de un clasificador lineal en el espacio de características; w define un hiperplano en el espacio de características que separa las proyecciones de datos de entrenamiento desde el origen.

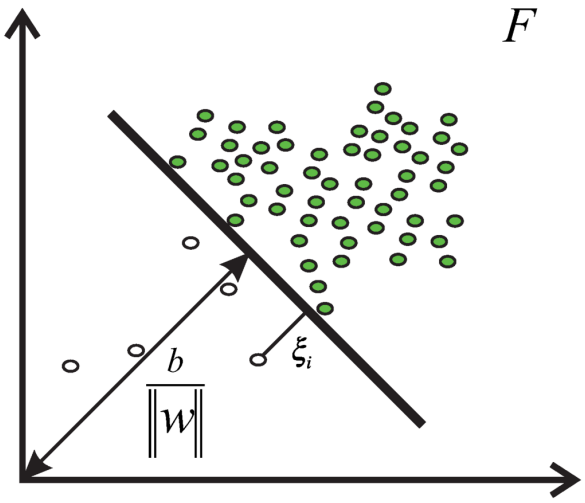


Figura 3.9: Hiperplano *One-Class SVM* en el espacio de características que separa las proyecciones de los datos de entrenamiento desde el origen.

Una función kernel positiva k, es definida como $k(x, x') = \langle \Phi(x), \Phi(x')\rangle$, que implícitamente mapea los datos de entrenamiento o de prueba x, en una espacio de

características de dimensiones superiores. Mediante la introducción de los multiplicadores de Lagrange α_i, la función de decisión en el espacio de datos de entrada X, viene dada por:

$$f(x) = sgn\left(\sum_{i=1}^{n} \alpha_i k(x_i, x) - \rho\right) \tag{3.21}$$

si $f(x) = -1$, la muestra de entrada x es clasificado como anormal; caso contrario, x es clasificado como normal.

3.2.3. Clasificación basada en la Distancia Mínima

La clasificación basada en la distancia mínima es una medida de similaridad entre los patrones desconocidos entrantes con respecto a los patrones o clases aprendidas en la fase de entrenamiento mediante el calculo de la distancia mínima entre estas. Sea x un vector de características de un patrón desconocido y sean $\{c_1, c_2, ..., c_n\}$ los n patrones entrenados, se calcula la distancia mínima entre el vector x con cada patrón entrenado c_i. El vector x es categorizado junto al patrón más cercano. En la Figura 3.10 se ilustra un esquema del clasificador de la distancia mínima.

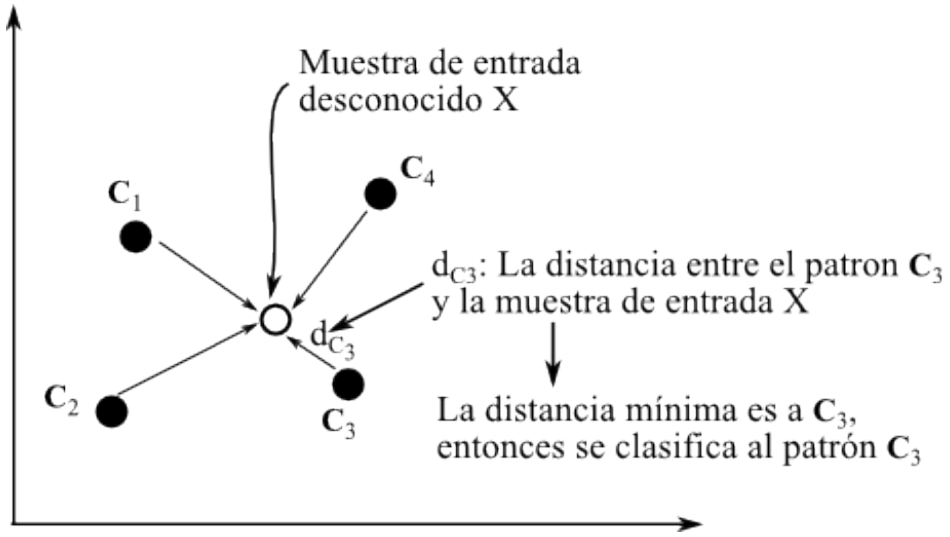


Figura 3.10: Esquema del clasificador de la distancia mínima.

La distancia se define como un índice de similitud de modo que la distancia mínima es idéntica a la similitud máxima. Las siguientes distancias a menudo se utilizan en este procedimiento.

3.2.3.1. Distancia Euclidiana

Se utiliza en los casos donde las varianzas de las clases entrenados son diferentes entre sí. La distancia Euclidiana es teóricamente idéntico al índice de similitud. Con la

ecuación siguiente se calcula la distancia mínima entre dos vectores:

$$d_i = \sqrt{(x - c_i)^T (x - c_i)} \tag{3.22}$$

3.2.3.2. Distancia de Mahalanobis

La distancia de Mahalanobis con la matriz de varianza-covarianza, es utilizado en los casos en que existe una correlación entre los ejes de espacio de características. La ecuación siguiente mide la distancia de Mahalanobis.

$$d_i = \sqrt{(x - c_i)^T \Sigma_i^{-1} (x - c_i)} \tag{3.23}$$

En las ecuaciones 3.22 y 3.23; $x = [x_1, x_2, ...x_k]$ representa un vector característico de k dimensiones, $c_i = [c_1, c_2, ..., c_k]$ es el i-ésimo patrón o clase entrenada y Σ_i es la matriz de varianza-covarianza.

3.3. Consideraciones Finales

En procesamiento de vídeo, los descriptores visuales cumplen un rol importante en la detección de eventos en vídeo, ya que son los encargados de describir los contenidos de los objetos y eventos presentes en un vídeo o imagen. En esta investigación se ha estudiado algunos de los descriptores más utilizados en la visión por computador. Por ejemplo, el enfoque de flujo óptico realiza la estimación de movimiento de manera óptima en vídeo, y para la detección de objetos basado en la apariencia se utilizan los histogramas de gradientes orientados (HOGs) que realizan un trabajo eficiente consiguiendo buenos resultados. Por otro lado, con el objetivo de clasificar la información extraída por los descriptores visuales, los algoritmos de clasificación son estudiados. Por ejemplo, el clasificador SVM es utilizado para abordar problemas de clasificación de dimensiones superiores. En cambio el clasificador de la distancia mínima es un clasificador básico que mide la similaridad entre dos patrones mediante el calculo de la distancia mínima.

Capítulo 4

Metodología Propuesta

En este capítulo se describe las etapas de desarrollo del modelo propuesto para la detección de eventos anómalos en secuencias de vídeo. El mismo consta de tres etapas principales, en la Figura 4.1 se muestra el esquema general. La primera etapa se denomina como la etapa de pre-procesamiento, donde filtros espaciales son utilizados directamente sobre los cuadros del vídeo con el objetivo de reducir las variaciones de intensidad y eliminar presencia de ruido. También en esta etapa se realiza la detección de regiones de movimiento (*foreground*) mediante la sustracción de cuadros consecutivos. La segunda etapa es la extracción de características, en esta etapa se realiza la extracción de los atributos de movimiento tales como la velocidad y la aceleración del movimiento, y la apariencia tales como las texturas del flujo óptico y el histograma de la gradiente del flujo óptico. Finalmente, la tercera etapa se denomina detección de eventos anómalos en secuencias de vídeo mediante el uso de la medida de distancia mínima. En las siguientes secciones se detallan con más profundidad cada una de las tres etapas.

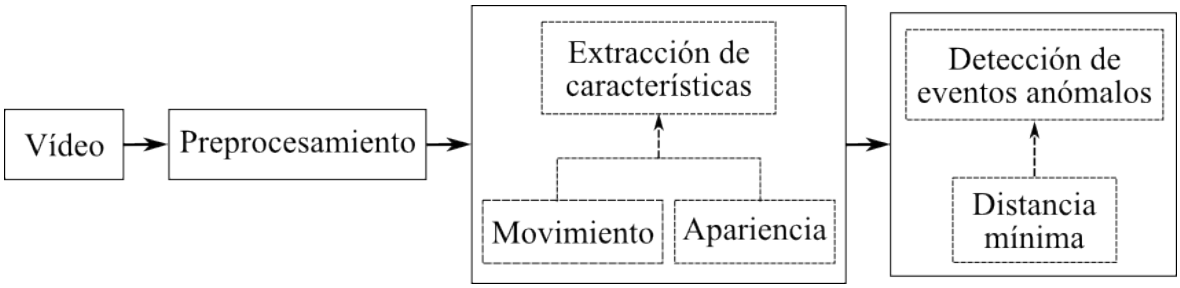


Figura 4.1: Esquema general del modelo propuesto.

4.1. Pre-procesamiento

El pre-procesamiento es una etapa fundamental en el procesamiento de imágenes y de vídeo, debido a que los vídeos o las imágenes de entrada muchas veces capturan información con presencia de valores atípicos. Por lo tanto, la primera etapa del modelo propuesto consiste en aplicar filtros espaciales a los cuadros del vídeo con el objetivo de reducir las variaciones de intensidad y eliminar la presencia de ruido. Existen muchos algoritmos en la literatura que realizan esta tarea de manera eficaz (por ejemplo, filtro Gaussiano, la mediana, bilateral, etc). En la presente investigación se utiliza el filtro Gaussiano para efecto de suavización y remoción del ruido presente en los cuadros. Además de realizar el suavizado, en esta etapa se realiza la detección de regiones en movimiento, mediante la sustracción de cuadros consecutivos del vídeo.

4.1.1. Filtro Gaussiano

Filtro Gausiano o conocido también como desenfoque Gaussiano es un efecto de suavizado que se aplica a las imágenes. En esencia, el efecto mezcla ligeramente las intensidades de los píxeles vecinos de una imagen con una máscara Gaussiana (calculado a partir de una función Gaussiana), lo que provoca que la imagen pierda algunos detalles minúsculos, de esta forma, hace que la imagen se vea más suave respecto a los bordes presentes en la imagen. En la Figura 4.2 se aplica el filtro Gaussiano a la primera imagen que contiene ruido y el resultado se observa en la segunda imagen.

(a) (b)

Figura 4.2: Suavizado Gaussiano de una imagen que contiene ruido (a) y el resultado del filtro Gaussiano se observa en (b).

4.1.2. Sustracción de Fondo

La sustracción de fondo es una etapa fundamental de una gran cantidad de aplicaciones en las que se necesita detectar movimiento, identificar y/o seguir objetos en secuencias de imágenes dinámicas. Entre estas aplicaciones se pueden nombrar a la vídeo-vigilancia, la detección y captura del flujo de movimiento, la interacción hombre-computadora o la codificación de vídeo basado en el contenido, entre otros. El proceso de sustracción es también denominado extracción de primer plano o de objetos en movimiento (*foreground extraction*), y consiste en una serie de métodos que permiten distinguir entre áreas de fondo o estáticas (*background*), y áreas dinámicas que corresponden al primer plano (*foreground*).

La sustracción de fondo es una técnica común y ampliamente utilizado para la generación de una máscara de primer plano. Es decir, una imagen binaria que contiene los píxeles pertenecientes a los objetos en movimiento en la escena. En esta investigación la sustracción de fondo se calcula mediante una resta entre dos cuadros consecutivos. Luego estableciendo un valor de umbral se genera la máscara de primer plano. La Figura 4.3 muestra el proceso de sustracción de fondo.

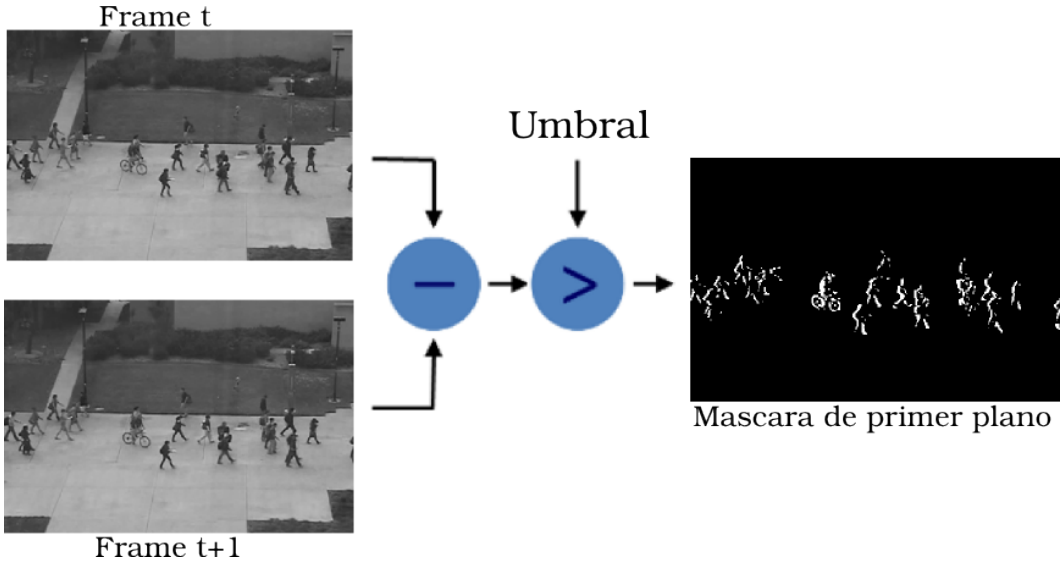


Figura 4.3: Proceso de detección de las regiones de movimiento.

4.2. Extracción de Características

La extracción de características en imágenes y secuencias de vídeo, consiste en extraer información contenida en la escena mediante descriptores visuales. Por lo general, se describen características elementales tales como la forma, el color, la textura o el movimiento, entre otros. Estos descriptores tienen que ser robustos, invariantes

a transformaciones geométricas, insensibles al ruido de captura y a los cambios de iluminación.

En esta tesis para la extracción de características, primero se procede a dividir la secuencia de vídeo en parches locales espacio-temporales o volúmenes 3D sin superposición. Luego, para cada parche espacio-temporal P, un vector característico es extraído conteniendo la información de movimiento, donde se calcula la velocidad y la aceleración del flujo (Subsección 4.2.1) y la información de apariencia, donde se calculan las texturas y el histograma de la gradiente del flujo óptico (Subsección 4.2.2), con el objetivo de detectar eventos anómalos en vídeo. Finalmente, se concatena ambos características en un vector característico final. La Figura 4.4 muestra un esquema del proceso de extracción de características.

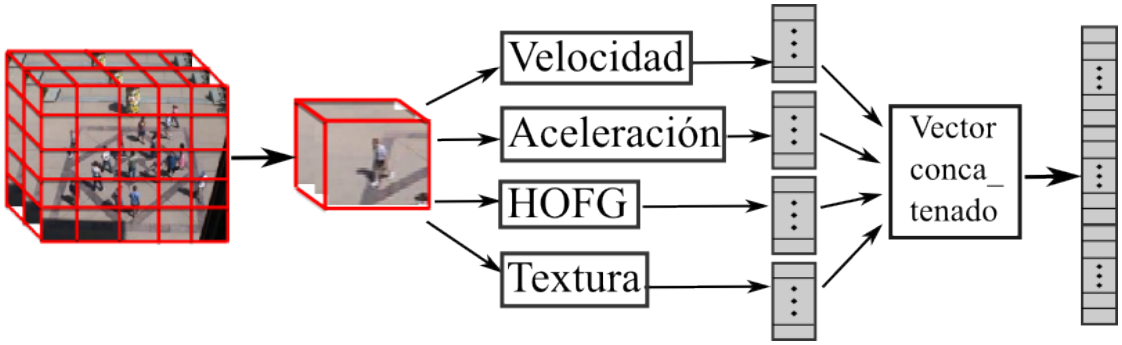


Figura 4.4: Esquema del proceso de extracción de características.

4.2.1. Característica de Movimiento

Para representar la característica de movimiento en la escena, se utiliza el algoritmo del flujo óptico de Lucas-Kanade piramidal (Sección 3.1.1). El flujo es calculado en cada píxel utilizando el método de Lucas-Kanade. Es importante resaltar que solo se utilizan los píxeles de primer plano (*foreground*) para representar la característica de movimiento. En esta tesis son usadas las características de velocidad (Ryan et al., 2011) y aceleración (Nallaivarothayan et al., 2014) del flujo óptico.

4.2.1.1. Velocidad

Para cada parche espacio-temporal P, la información de velocidad del flujo óptico se calcula directamente realizando la sumatoria del desplazamiento de los componentes del flujo óptico. Se crea un vector de características de dos dimensiones que contiene tanto los componentes horizontal y vertical de la velocidad:

$$V_u = \sum_{(x,y)\in P} u\,(x,y)\,, \quad V_v = \sum_{(x,y)\in P} v\,(x,y)$$

donde (x, y) son las coordenadas de todos los píxeles que pertenecen al parche P, (u, v) son los componentes horizontal y vertical del flujo óptico y $[V_u, V_v]$ representa la información de velocidad del flujo óptico, estas características ayudan a modelar la velocidad esperada en cada región de la escena.

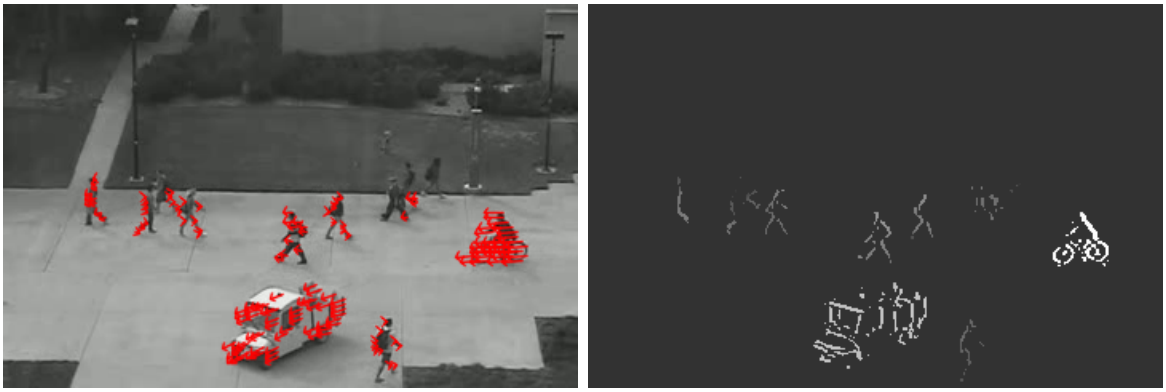
4.2.1.2. Aceleración

La característica de aceleración del flujo óptico extrae información sobre la variación del flujo óptico temporalmente (Nallaivarothayan et al., 2014). Es decir, se puede observar que el flujo óptico del cuerpo humano varía en el tiempo debido a la naturaleza de los movimientos de sus extremidades, en particular debido al movimiento de las piernas. Esto da lugar a que la aceleración del cuerpo humano varíe de manera significativa en la dirección, pero con una pequeña magnitud. Además, objetos como vehículos y bicicletas tienden a tener una aceleración alta debido al impulso aplicado a ellos, pero la dirección de su aceleración es predominantemente uniforme debido a su movimiento rígido.

Para modelar la información de aceleración del flujo. En primer lugar, se calcula los vectores del flujo óptico a nivel del píxel para cada cuadro del vídeo utilizando el método de Lucas-Kanade piramidal (Bouguet, 2001). Luego, se calculan las magnitudes del flujo óptico para cada cuadro y se normaliza estas magnitudes entre $[0, 255]$ para crear una imagen como se muestra en la Figura 4.5 (b), donde la magnitud del flujo óptico en la posición (x, y) esta dado por:

$$M(x, y) = \sqrt{u^2 + v^2}$$

donde u y v son los componentes horizontal y vertical del vector del flujo óptico, respectivamente. La información de aceleración en cada píxel se calcula a partir de la derivada de tiempo de la imagen de magnitud del flujo óptico. Es decir, se calcula los



(a) (b)

Figura 4.5: La imagen del campo del flujo óptico se muestra en (a) y la imagen normalizado entre $[0, 255]$ de la magnitud del flujo óptico en (b).

vectores del flujo óptico que existe entre dos imágenes de magnitud del flujo óptico consecutivos. Finalmente, la información de aceleración para un parche P se calcula directamente utilizando una sumatoria de los vectores de la aceleración en nivel del píxel para crear un vector de características de dos dimensiones que contiene tanto los componentes horizontal y vertical de la aceleración $[A_u, A_v]$,

$$A_u = \sum_{(x,y)\in P} u^{'}(x,y), \quad A_v = \sum_{(x,y)\in P} v^{'}(x,y)$$

donde (x, y) son las coordenadas de todos los píxeles que pertenecen al parche P, $(u^{'}, v^{'})$ son los componentes horizontal y vertical del flujo óptico de la imágenes de magnitud.

4.2.2. Característica de Apariencia

Debido a la naturaleza no rígida del movimiento del cuerpo humano causada por los movimientos de las extremidades, el flujo óptico varía a través del cuerpo humano, creando así una necesidad de extraer información acerca de las variaciones espaciales en el flujo óptico. Al mismo tiempo, los objetos anormales tales como bicicletas y furgonetas tienen flujo uniforme y suave a través de su superficie. Así la información adicional acerca de la variación de flujo a través de la superficie de un objeto puede ser indicativa de las variaciones de la apariencia, además de proporcionar la información de movimiento. Esto puede ayudar en la detección de objetos anormales tales como bicicletas con movimiento lento que no presentan variación significativa de movimiento para clasificarlos como anomalías.

Para abordar los problemas anteriores, en esta tesis se utiliza como características de apariencia, el histograma de gradientes del flujo óptico (Nallaivarothayan et al., 2014) y la textura del flujo óptico que mide la uniformidad del movimiento (Ryan et al., 2011).

4.2.2.1. Histograma de Gradientes de Flujo Óptico

El Histograma de Gradientes del Flujo Óptico (HOFG del inglés *Histogram of Optical Flow Gradients*) mide la variación espacial de movimiento (Nallaivarothayan et al., 2014). Donde se representa a los dos componentes de flujo óptico horizontal y vertical como imágenes separadas y se calcula los gradientes de cada imagen utilizando los operadores de Sobel. Un histograma de cuatro *bins* es generado para cada uno de los componentes del flujo horizontal y vertical, y un enfoque basado en *binning* suavizado es utilizado para calcular el peso asignado a dos *bins* adyacentes basado en la distancia a los centros del *bin*. Además, estos votos suavizados son ponderados basados en la magnitud del gradiente.

Los componentes del gradiente del flujo horizontal en la posición del píxel p son dadas por $u_x(p) = \frac{\delta}{\delta x}(u)$ y $u_y(p) = \frac{\delta}{\delta y}(u)$, donde u_x y u_y son los componentes de

gradiente horizontal y vertical de flujo horizontal. Del mismo modo, los componentes del gradiente del flujo vertical son dadas por $v_x(p) = \frac{\delta}{\delta x}(v)$ y $v_y(p) = \frac{\delta}{\delta y}(v)$, donde v_x y v_y son los componentes de gradiente horizontal y vertical de flujo vertical. La orientación del gradiente de flujo óptico horizontal en la posición de píxel p es dada por $\theta_h(p) = \arctan\frac{u_y(p)}{u_x(p)}$ y la orientación del flujo vertical, $\theta_v(p)$ se puede calcular de manera similar.

Después de calcular la orientación en cada píxel, los pesos se asignan a los *bins* basado en la orientación calculado. No se considera el signo de la orientación como las técnicas utilizadas en la detección humana, y para cualquier ángulo negativo se agrega π radianes para convertirlo a su contraparte ángulo positivo. Se utiliza cuatro *bins* centradas en los ángulos: $\frac{\pi}{8}$, $\frac{3\pi}{8}$, $\frac{5\pi}{8}$ y $\frac{7\pi}{8}$. Los pesos se asignan a los *bins* siguiendo un proceso de suavización, por lo tanto, para la orientación del gradiente $\theta_h(p)$, en la posición p, que se encuentra entre dos centros θ_n y θ_{n+1} del *bin* adyacente; es decir $\theta_n < \theta_h(p) < \theta_{n+1}$, el peso suavizado para el *bin* n-ésimo está dada por:

$$h_n^h(p) = \frac{\theta_{n+1} - \theta_h(p)}{\theta_{n+1} - \theta_n}\sqrt{u_x^2(p) + u_y^2(p)} \tag{4.1}$$

y el peso para $(n+1)$-ésimo *bin* está dada por:

$$h_{n+1}^h(p) = \frac{\theta_h(p) - \theta_n}{\theta_{n+1} - \theta_n}\sqrt{u_x^2(p) + u_y^2(p)}. \tag{4.2}$$

Todos los demás *bins* se mantienen sin cambios para el gradiente en particular. Los pesos del histograma para los gradientes de la componente vertical del flujo óptico se calculan de una manera similar.

Finalmente, la información del histograma para un parche de espacio-temporal P, se incorpora directamente utilizando una suma de los valores del histograma a nivel del píxel, el n-ésimo *bin* del gradiente de flujo óptico horizontal y el gradiente de flujo óptico vertical, están dadas por $H_n^h = \sum_{p\in P} h_n^h(p)$ y $H_n^v = \sum_{p\in P} h_n^v(p)$, respectivamente.

Los dos histogramas de un parche se concatenan en un solo vector de características de ocho dimensiones,

$$\left[H_{\frac{\pi}{8}}^h, H_{\frac{3\pi}{8}}^h, H_{\frac{5\pi}{8}}^h, H_{\frac{7\pi}{8}}^h, H_{\frac{\pi}{8}}^v, H_{\frac{3\pi}{8}}^v, H_{\frac{5\pi}{8}}^v, H_{\frac{7\pi}{8}}^v\right].$$

4.2.2.2. Texturas del Flujo Óptico

La textura del flujo óptico mide la uniformidad del movimiento de la escena, se calcula a partir del producto escalar de los vectores del flujo óptico en diferentes desplazamientos (*offsets*). La uniformidad del movimiento calculado en diferentes desplazamientos es útil para la detección de objetos de diferentes tamaños (Ryan et al., 2011).

Para calcular la información de textura de un parche espacio-temporal P. Se calcula el producto escalar de los vectores del flujo óptico de los píxeles p y $p' = p + \delta$,

donde δ denota el desplazamiento del píxel p, mediante,

$$\phi_\delta = \sum_{p \in P} [u(p)u(p+\delta) + v(p)v(p+\delta)]$$

donde u y v representan los componentes del flujo óptico horizontal y vertical respectivamente. En esta tesis se utiliza múltiples valores de desplazamientos ($\delta = 1, 3, 5$), para lograr un análisis multi-escala. El vector característico final esta representado por:

$$[\phi_{(1,1,0)}, \phi_{(3,3,0)}, \phi_{(5,5,0)}]$$

donde se captura la uniformidad en desplazamientos de 1, 3 y 5 píxeles.

4.3. Detección de Eventos Anómalos

En las aplicaciones reales de detección de eventos anómalos por lo general predominan las muestras de eventos normales, y las muestras de eventos anómalos tienen una presencia escasa, debido a esto la etapa de entrenamiento de un modelo se hace aún más difícil. Para abordad este problema, en esta etapa se plantea el uso del clasificador de la distancia mínima para detectar eventos anómalos.

4.3.1. Clasificación basada en la Distancia Mínima

La clasificación basada en la distancia mínima es una técnica que mide la semejanza entre dos muestras (Sección 3.2.3). La idea principal de esta forma de clasificación es buscar alguna muestra entrenada que sea similar a la muestra de entrada (Colque et al., 2015). La Figura 4.6 ilustra este paso de clasificación, donde los puntos azules representan a las muestras entrenadas y los puntos anaranjados a las muestras entrantes. Para determinar si una muestra entrante es anómalo se establece un umbral U_t. Si una muestra entrante es similar a alguna muestra conocida (es decir, es menor que el umbral U_t), entonces esta muestra es clasificado como normal (caso del punto B); caso contrario, si la muestra entrante no es similar a ninguna muestra entrenada, entonces esta muestra será clasificado como anómalo (caso del punto A).

La clasificación se realiza a nivel de cuadro (*frame-level*), para determinar si un cuadro es normal se verifica que todos los parches espacio-temporales dentro del cuadro son clasificados normales, caso contrario es considerado anormal.

4.3.1.1. Problemas de Perspectiva

El clasificador de distancia mínima presenta problemas de detección en vídeos que presentan problemas de perspectiva, esto debido a la distorsión de la perspectiva en una

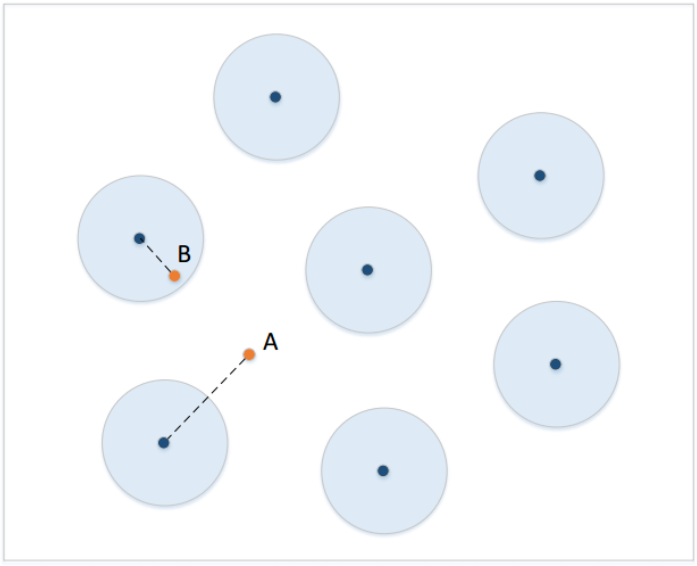


Figura 4.6: Proceso de clasificación. Una muestra de evento anómalo es representado por el punto A y una muestra de evento normal por el punto B. Originalmente mostrado en (Colque et al., 2015).

escena, los objetos cercanos a la cámara parecen ser grandes mientras que los objetos alejados a la cámara parecen ser pequeños. Esto puede afectar significativamente a los métodos de extracción de características como también a los clasificadores, tal es el caso de la distancia mínima. La Figura 4.7 muestra la variación del flujo óptico del mismo carro en dos diferentes profundidades de la escena.

(a) Un carro más cerca de la cámara (b) Un carro más lejos de la cámara

Figura 4.7: Vídeos con problema de perspectiva, se observa la variación del flujo óptico según la profundidad de la escena.

Para abordar estos problemas y obtener mejores resultados, se propone una clasificación por regiones locales. Es decir, la escena se divide en regiones espaciales locales de acuerdo a la profundidad de la escena, donde en cada región son modeladas sus respectivas muestras que representan eventos normales. En esta tesis se ha utilizado

dos tipos de clasificación, 01 y 04 regiones locales, como se muestra en la Figura 4.8. Para cada región las etapas de entrenamiento y prueba se realizan por separado.

(a) Una (01) región local (b) Cuatro (04) regiones locales

Figura 4.8: Dos tipos de clasificación por regiones locales propuesta en esta tesis.

4.4. Consideraciones Finales

El modelo propuesto trabaja en contextos predefinidos (en este caso áreas estrictamente peatonales) y se utiliza la misma metodología para detectar los eventos anómalos tanto globales y locales. También en este modelo se plantea la combinación de las informaciones de movimiento y apariencia con el objetivo de superar los resultados de los métodos de la literatura. Además, se propone una clasificaciones por regiones locales para abordar los vídeos que presentan problemas de perspectiva.

Capítulo 5

Resultados Experimentales

En este capítulo, se presenta los resultados y experimentos de la investigación. Muchos experimentos se han llevado acabo con el fin de evaluar la eficacia y la robustez del modelo propuesto sobre los bases de datos (*datasets*) que están disponibles públicamente. En primer lugar, en la Sección 5.1 se explica a detalle los bases de datos que se han utilizado para evaluar el algoritmo propuesto. Luego, los criterios y los parámetros generales para la evaluación del algoritmo se definen en la Sección 5.2. Finalmente, en la Sección 5.3 se analiza los resultados y se compara el modelo propuesto con los enfoques existentes en la literatura.

5.1. Base de Datos

Para evaluar la robustez y el rendimiento del algoritmo propuesto para la detección de eventos anómalos, se ha utilizado dos bases de datos públicas. La base de datos de UMN *Crowd Dataset* (UMN, 2012) y la base de datos de UCSD *Anomaly Detection Dataset* (UCSD, 2013). La primera base de datos ha sido utilizado para evaluar la detección del Evento Anómalo Global (EAG) y la segunda base de datos ha sido utilizado para la detección del Evento Anómalo Local (EAL).

5.1.1. Base de Datos de UMN

La base de datos de UMN (UMN, 2012) consiste de tres diferentes escenarios de eventos de escape de multitud de personas, dos escenarios exteriores (*outdoor*) y un escenario en ambiente cerrado (*indoor*); con una resolución de 320×240 píxeles. Los eventos son considerados normales cuando los peatones caminan al azar en diferentes direcciones en una plaza o en un centro comercial, y los eventos anómalos cuando los peatones se dispersan (huyen) al mismo tiempo. Hay un total de 11 vídeos en toda la

base de datos, cada vídeo consiste en una parte inicial de evento normal y termina con secuencias de evento anormal. La Figura 5.1 muestra los cuadros de los tres escenarios.

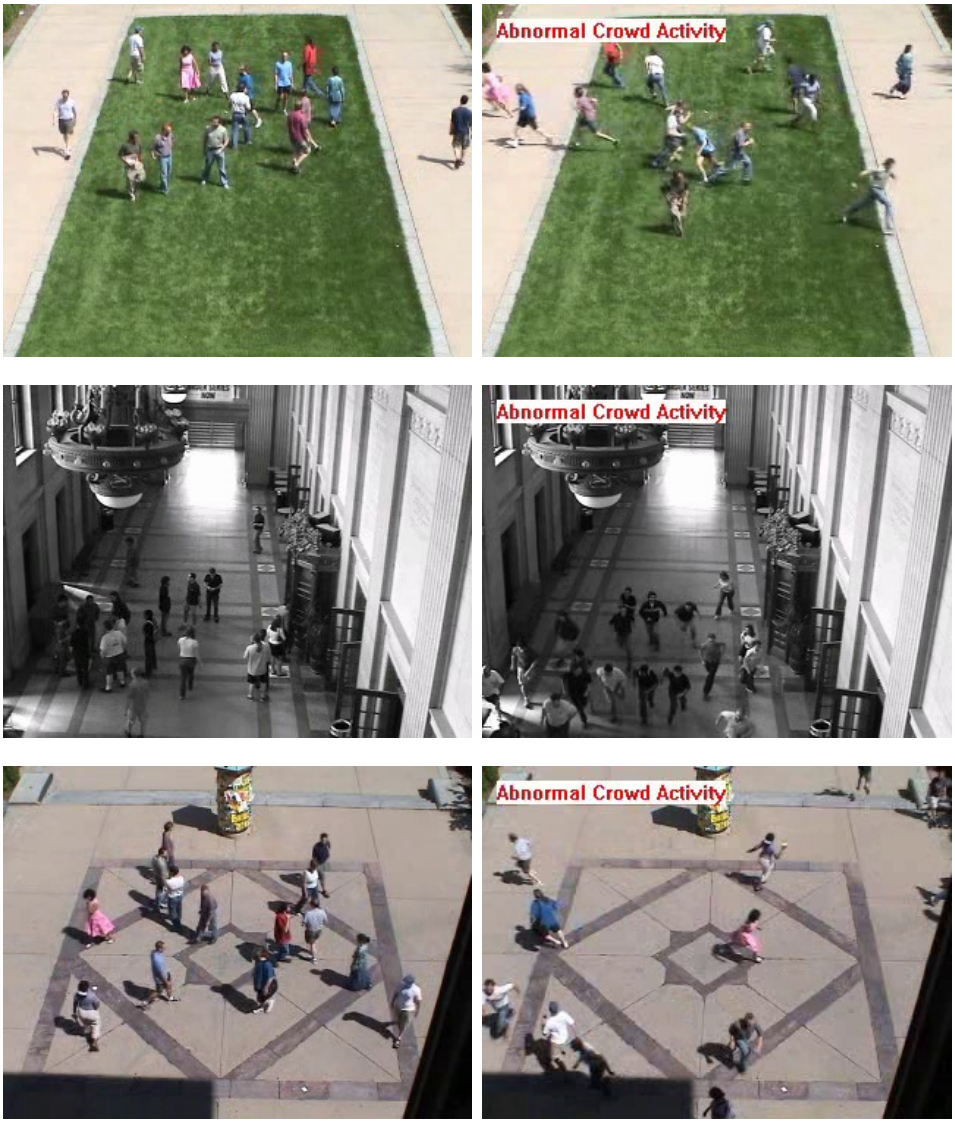


Figura 5.1: Cuadros de los tres escenarios de la base de datos de UMN, primera fila muestra los cuadros del escenario 1, segunda fila los cuadros del escenario 2 y la tercera fila del escenario 3.

5.1.2. Base de Datos de UCSD

La base de datos de UCSD (UCSD, 2013), ha sido adquirido con una cámara estática montada en altura, proyectado a una zona peatonal. Esta base de datos esta dividido en dos grupos de bases de datos Peds1 y Peds2, cada uno contiene vídeos de

dos diferentes escenarios. Los eventos anómalos incluyen carros, bicicletas, motocicletas y patinadores. Los autores de este base de datos también clasifican sillas de ruedas y coches de bebé como anomalías. Aunque estas entidades no se pueden considerar como eventos anómalos en zonas peatonales, en esta tesis se consideran como anomalías para una comparación exacta con los métodos en la literatura.

5.1.2.1. UCSD Peds1

La base de datos UCSD Peds1 contiene un total de 70 vídeos, 34 vídeos de entrenamiento que consiste solo de cuadros normales y 36 vídeos de prueba que consiste de ambos cuadros normales y anormales. Cada vídeo contiene 200 cuadros con una resolución de 258×158 píxeles. Un cuadro normal y anormal de esta base de datos se muestra en la Figura 5.2.

(a) (b)

Figura 5.2: Cuadros de la base de datos UCSD Peds1, (a) cuadro normal y (b) cuadro anormal.

5.1.2.2. UCSD Peds2

La base de datos UCSD Peds2 contiene un total de 28 vídeos, 16 vídeos de entrenamiento que consiste solo de cuadros normales y 12 vídeos de prueba que consiste de ambos cuadros normales y anormales. Los vídeos contienen entre 120 y 180 cuadros con una resolución de 360×240 píxeles. Un cuadro normal y anormal de esta base de datos se muestra en la Figura 5.3.

(a) (b)

Figura 5.3: Cuadros de la base de datos UCSD Peds2, (a) cuadro normal y (b) cuadro anormal.

5.2. Criterios de Evaluación y Parámetros

El criterio de evaluación que se ha aplicado en este modelo propuesto es a nivel de cuadro (*frame-level*). Este criterio verifica si un cuadro contiene al menos un (01) parche anómalo, entonces este es considerado como detección. Un umbral L_T es utilizado para detectar los parches espacio-temporales anómalos. Estas detecciones se comparan con la anotación del *ground truth* de cada cuadro. Es importante resaltar que esta evaluación no comprueba si la detección coincide con la ubicación real del evento anómalo.

Para medir el rendimiento del método propuesto, la curva ROC o la curva de la Característica Operativa del Receptor (ROC, del inglés *Receiver Operating Characteristic*), el Área Bajo la Curva (AUC, del inglés *Area Under Curve*) y la Tasa de Error Igual (EER, del inglés *Equal Error Rate*) son calculados. La curva ROC esta formado por la Tasa de Verdaderos Positivos (TPR, del inglés *True Positive Rate*) y la Tasa de Falsos Positivos (FPR, del inglés *False Positive Rate*), de las cuales la TPR determina un clasificador mediante la clasificación de casos positivos correctamente entre todas las muestras positivas disponibles durante la prueba y la FPR determina cuantos resultados positivos incorrectos se producen entre todas las muestras negativas disponibles durante la prueba. Estas medidas están dadas por las siguientes ecuaciones:

$$TPR = \frac{verdaderospositivos}{verdaderospositivos + falsosnegativos},$$

$$FPR = \frac{falsospositivos}{falsospositivos + verdaderosnegativos},$$

donde los verdaderos positivos son los eventos anómalos correctamente etiquetados, falsos negativos son los eventos normales incorrectamente etiquetados, falsos positivos son los eventos anómalos incorrectamente etiquetados y los verdaderos negativos son los eventos normales correctamente etiquetados. Además, el valor de umbral L_T es variado para generar la curva ROC.

El AUC es el área bajo la Curva ROC generado, también conocido matemáticamente como integral definida. En esta tesis se utilizó la siguiente ecuación para calcular el AUC.

$$AUC = \sum_{i=1}^{n} \left[\frac{(TPR_i + TPR_{i-1})\,(FPR_i + FPR_{i-1})}{2} \right] \tag{5.1}$$

donde $i = 1, ..., n$ y n es el número de pruebas que se realizó al algoritmo propuesto con umbral L_T diferente.

La Tasa de Error Igual (EER) se calcula a partir de la Tasa de Falsos Positivos (FPR) y la Tasa de Falsos Negativos (FNR, del inglés *False Negative Rate*). Cuando las tasas son iguales, el valor común se conoce como la Tasa de Error Igual o EER. El código fuente en Matlab para calcular la EER se muestra a continuación.

```
% the_fpr_fnr es una matriz de 2xn conformado por fpr como fila 1 y fnr
    como fila 2.
% fpr: Tasa de Falsos Positivos
% fnr: Tasa de Falsos Negativos
diferencia = diff(the_fpr_fnr(1:2,:),[],1);
indice = find(abs(diff(sign(diferencia))) ~= 0);
valores = the_fpr_fnr(1:2,indice:indice+1);
eer = mean(valores(:),1); % El valor de EER
```

Por otro lado, el modelo propuesto requiere de algunos parámetros y configuraciones generales para evaluar el rendimiento del algoritmo propuesto:

- En la fase extracción de características, se divide cada cuadro entrante en parches espacio-temporales sin superposición. La dimensión de los parches espacio-temporales se ha fijado de 20×20 píxeles espacialmente y 7 cuadros temporalmente, para todas las bases de datos. Luego, para cada parche se genera un vector característico que contiene las características explicadas en la Sección 4.2. Es decir, el vector característico contiene dos características que describen la velocidad del flujo óptico horizontal y vertical, dos características de la aceleración del flujo óptico en las direcciones horizontal y vertical, un histograma de ocho dimensiones de los componentes de la gradiente del flujo óptico (un histograma de cuatro dimensiones para cada dirección horizontal y vertical) y tres características que describen la textura del flujo óptico. Por lo tanto, el vector característico final para cada parche espacio-temporal es de 15 dimensiones. Finalmente, la detección se realiza mediante la técnica de clasificación explicado en la Sección 4.3.
- Para reducir al mínimo las falsas alarmas, se ha implementado un filtraje de post-procesamiento temporal como proceso final de clasificación. Si un parche en el tiempo t se clasificó inicialmente como anómalo, verificamos sus vecinos temporales inmediatos. Es decir, si al menos tres vecinos en el tiempo $(t-1, t-2, t-3)$ se clasificaron como anómalos, se asume que el parche fue clasificado correctamente como anómala. De lo contrario, se reclasifica como parche normal (es decir, no anómalo). En esta tesis la dimensión del post-procesamiento se ha establecido en $t = 5$.

5.3. Resultados de la Detección de Eventos Anómalos

En esta sección, se presentan los resultados del método propuesto de detección de eventos anómalos. La sección se divide en dos subsecciones, en la Subsección 5.3.1 se presenta los resultados de la detección del evento anómalo global (EAG) utilizando la base de datos de UMN y en la Subsección 5.3.2 se presenta los resultados de la detección del evento anómalo local (EAL) utilizando la base de datos de UCSD.

5.3.1. Detección del Evento Anómalo Global

La base de datos de UMN ha sido utilizado para evaluar la detección de EAG. Para la fase de entrenamiento se ha utilizado los primeros 300 cuadros de cada vídeo y los cuadros restantes se utilizaron en la fase de prueba.

Los resultados del método propuesto sobre la base de datos UMN se muestran en la Figura 5.4, donde se ilustra que la primera, segunda y tercera fila de imágenes corresponde a los resultados de la escena 1, 2 y 3 de la base de datos UMN, respectivamente. Además, se puede observar los dos tipos de clasificación propuestas en esta tesis (una región y cuatro regiones) en la primera y segunda columna, respectivamente. Cabe resaltar que los rectángulos en color rojo representan los parches espacio-temporales detectados por la clasificación propuesta.

En la Tabla 5.1 se muestra los resultados cuantitativos de nuestro experimento sobre la base de datos UMN, utilizando los dos tipos de clasificación propuestas (01 y 04 regiones). Para realizar el análisis del rendimiento del método propuesto se calcula el área bajo la curva (AUC) para cada escena. Los resultados sobre las escenas 1 y 3, utilizando la clasificación con una sola región logra un AUC de 0.9985 y 0.9954 respectivamente. Estos resultados superan a la clasificación con 04 regiones. Mientras tanto, en la escena 2 los resultados utilizando 04 regiones de clasificación logra un AUC 0.9486, el cual supera al resultado de la clasificación utilizando una sola región. Cabe resaltar que los vídeos de la escena 2 presentan problemas de perspectiva (ver Sección 4.3). Por lo tanto, se puede observar que la clasificación de la distancia mínima

Método propuesto	AUC (01 región)	AUC (04 regiones)
Escena 1	**0.9985**	0.9962
Escena 2	0.9182	**0.9486**
Escena 3	**0.9954**	0.9951

Tabla 5.1: Resultados cuantitativos del método propuesto sobre la base de datos UMN utilizando los dos tipos de clasificación propuestas (01 y 04 regiones). El área bajo la curva (AUC) de la curva ROC es calculado.

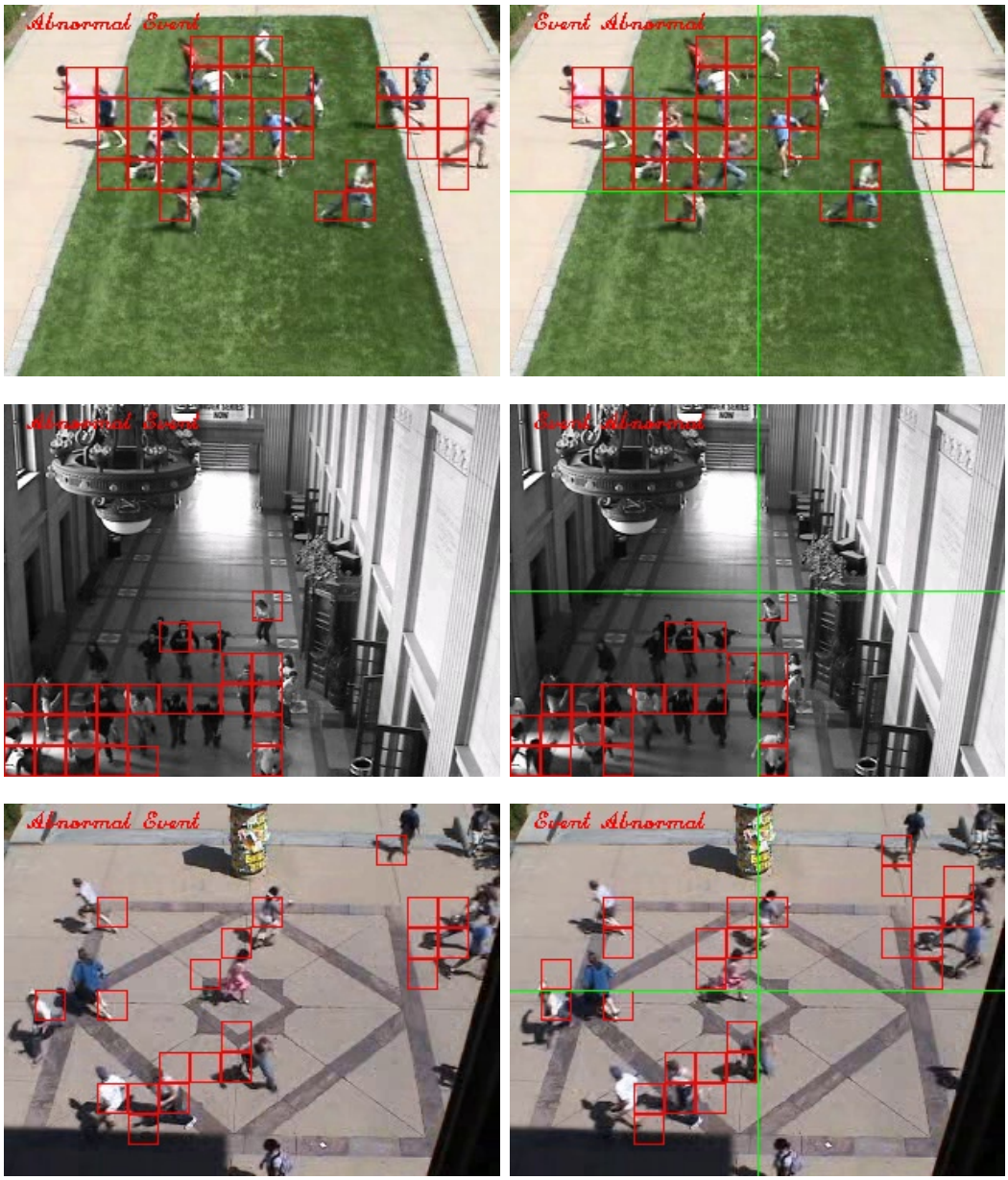


Figura 5.4: Resultados del método propuesto sobre la base de datos UMN. La primera, segunda y tercera fila de imágenes corresponde a las escenas 1, 2 y 3 de la base de datos UMN, respectivamente. La primera y la segunda columna de imágenes representa los resultados utilizando los dos tipos de clasificación propuestos, 01 y 04 regiones, respectivamente.

utilizando 04 regiones mejora los resultados en los vídeos que presentan problemas de perspectiva.

En la Figura 5.5 se muestra las curvas ROC del método propuesto utilizando los dos tipos de clasificación propuestas (01 y 04 regiones) sobre la base de datos UMN. Como se puede observar el rendimiento del método propuesto en la escena 2 utilizando

04 regiones de clasificación supera a la clasificación con una sola región, y en las escenas 1 y 3 el rendimiento es similar con los dos tipos de clasificación propuestas.

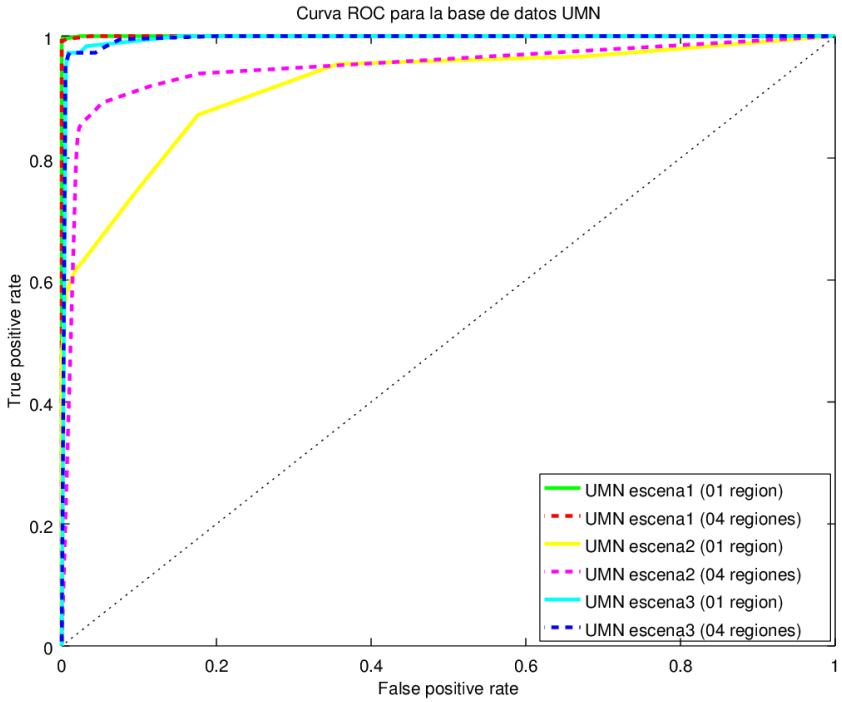


Figura 5.5: Las curvas ROC para la detección de EAG sobre la base de datos UMN, utilizando los dos tipos de clasificación (01 y 04 regiones).

5.3.1.1. Comparación con otros Métodos de la Literatura

La Tabla 5.2 se muestra la comparación cuantitativa del método propuesto con los métodos de la literatura tales como: el modelo de fuerza social (Mehran et al., 2009), el modelo de invariante caótica (Wu et al., 2010), el modelo de mezcla Gaussiana espacio-temporal (Shi et al., 2010), el método de costo de reconstrucción disperso (Cong et al., 2013). Se puede observar que el método propuesto supera a los resultados de la literatura en las escenas 1 y 3 obteniendo 0.998 y 0.995 de AUC, respectivamente. Sin embargo, los resultados del método propuesto en la escena 2 es comparable con los resultados de la literatura, debido a que los vídeos de esta escena presentan problemas de perspectiva. Cabe resaltar que el criterio de evaluación de los primeros dos modelos mencionados (Mehran et al., 2009; Wu et al., 2010) es diferente al criterio de evaluación de este modelo propuesto. Estos métodos realizan la evaluación considerando todas las escenas como un solo conjunto de vídeos de esta base de datos, obteniendo así un único resultado.

Métodos	AUC
Fuerza social (Mehran et al., 2009)	0.960
Invariante caótica (Wu et al., 2010)	0.990
Escena1 (Shi et al., 2010)	0.936
Escena2 (Shi et al., 2010)	0.775
Escena3 (Shi et al., 2010)	0.966
Esparso Escena1 (Cong et al., 2013)	0.995
Esparso Escena2 (Cong et al., 2013)	**0.975**
Esparso Escena3 (Cong et al., 2013)	0.964
Propuesta Escena1	**0.998**
Propuesta Escena2	0.948
Propuesta Escena3	**0.995**

Tabla 5.2: La comparación del método propuesto con los métodos de la literatura sobre la base de datos de UMN, se puede observar que nuestro método propuesto supera en las escenas que no presentan problemas de perspectiva. El área bajo la curva (AUC) de la curva ROC es calculado.

5.3.2. Detección del Evento Anómalo Local

Para la detección del EAL se realizó el experimento sobre las base de datos UCSD. En la Subsección 5.1.2 se describe la distribución de los vídeos para la fase de entrenamiento y de prueba. Esta base de datos contiene dos conjuntos de vídeos Peds1 y Peds2, cada uno ha sido grabado en un escenario diferente. Los experimentos se han realizado por separado para cada conjunto de vídeos Peds1 y Peds2, respectivamente.

Los resultados del experimento sobre la base de datos UCSD Peds1 y Peds2 se muestran en la Figura 5.6 y 5.7 respectivamente, en cada fila de imágenes se ilustra la detección de los diferentes objetos anómalos tales como: carros, bicicletas y patinadores. Además, cada columna representa los resultados de la detección utilizando los dos tipos de clasificación del método propuesto. La primera columna representa los resultados de clasificación tomando los cuadros como una sola región y la segunda columna dividiendo en cuatro regiones de clasificación.

La Tabla 5.3 muestra los resultados cuantitativos del experimento sobre la base de datos UCSD Peds1 y Peds2, utilizando los dos tipos de clasificación propuestos, primero tomando cada cuadro como una sola región y luego utilizando cuatro regiones de clasificación. Los resultados sobre la base de datos UCSD Peds1 logra un EER de 29.28 % y un AUC de 0.7923 utilizando cuatro regiones de clasificación, este resultado supera a la clasificación utilizando solo una región, esto ocurre debido a que los vídeos de esta base de datos presentan problemas de perspectiva (Sección 4.3). Cuando un vídeo presenta el problema de perspectiva esto afecta la extracción de características y en consecuencia dificulta la clasificación. Para superar este problema se propone la

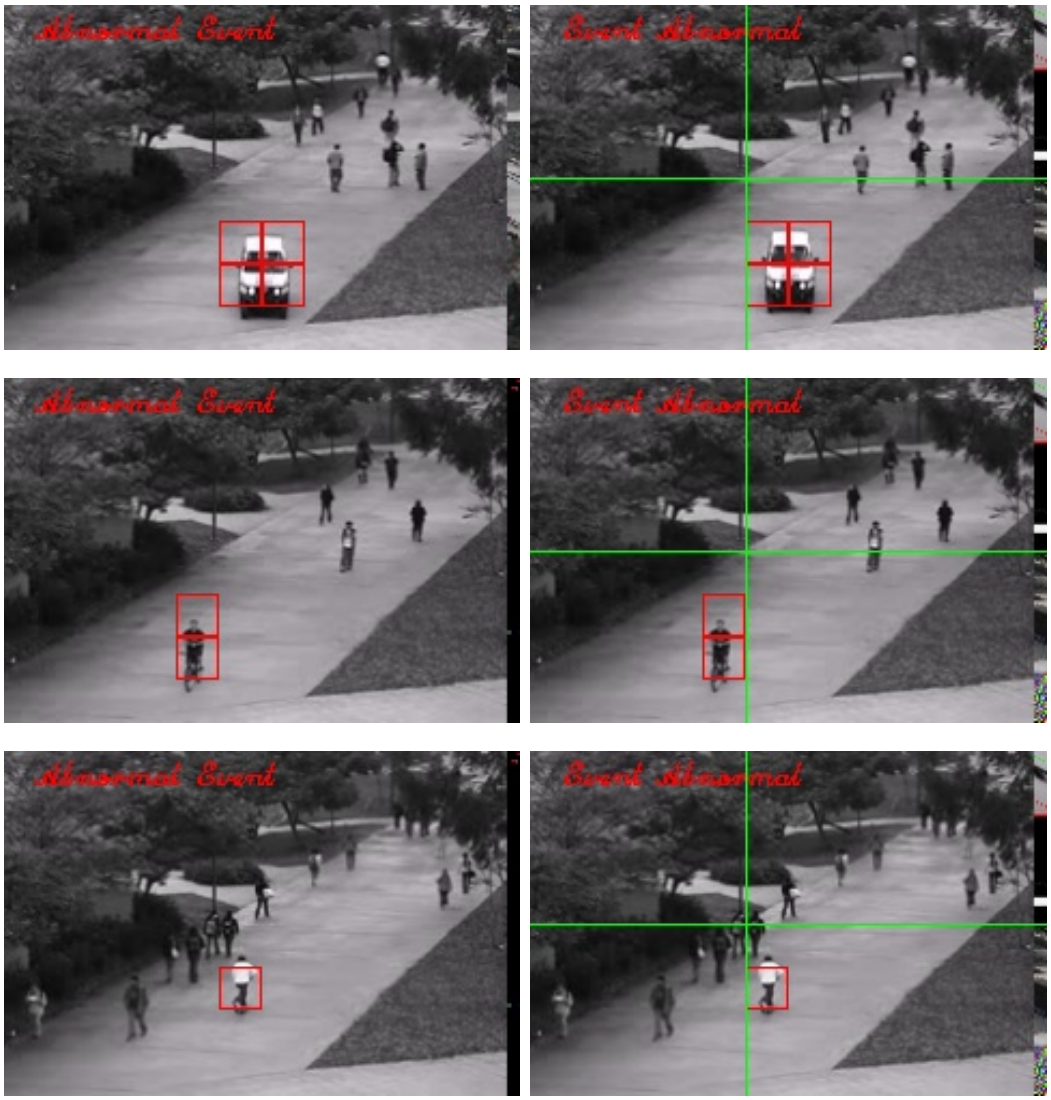


Figura 5.6: Resultados del método propuesto sobre la base de datos UCSD Peds1, en cada fila de imágenes se muestra la detección de los diferentes objetos como: carros, bicicletas y patinadores. La primera y la segunda columna de imágenes representa los resultados utilizando los dos tipos de clasificación propuestas, 01 y 04 regiones, respectivamente.

clasificación por regiones (en este caso se usó cuatro regiones), consiguiendo mejorar el rendimiento del método propuesto como se puede observar en el tabla de resultados. Por otro lado, los resultados sobre la base de datos Peds2 logra un EER de 07.24 % y un AUC de 0.9778 utilizando una sola región de clasificación, este resultado supera a la clasificación utilizando cuatro regiones, esto es debido a que los vídeos de esta base de datos no presentan problemas de perspectiva.

Por lo tanto, analizando los resultados obtenidos podemos concluir que la clasificación por regiones (cuatro regiones) es más adecuado para superar los vídeos que

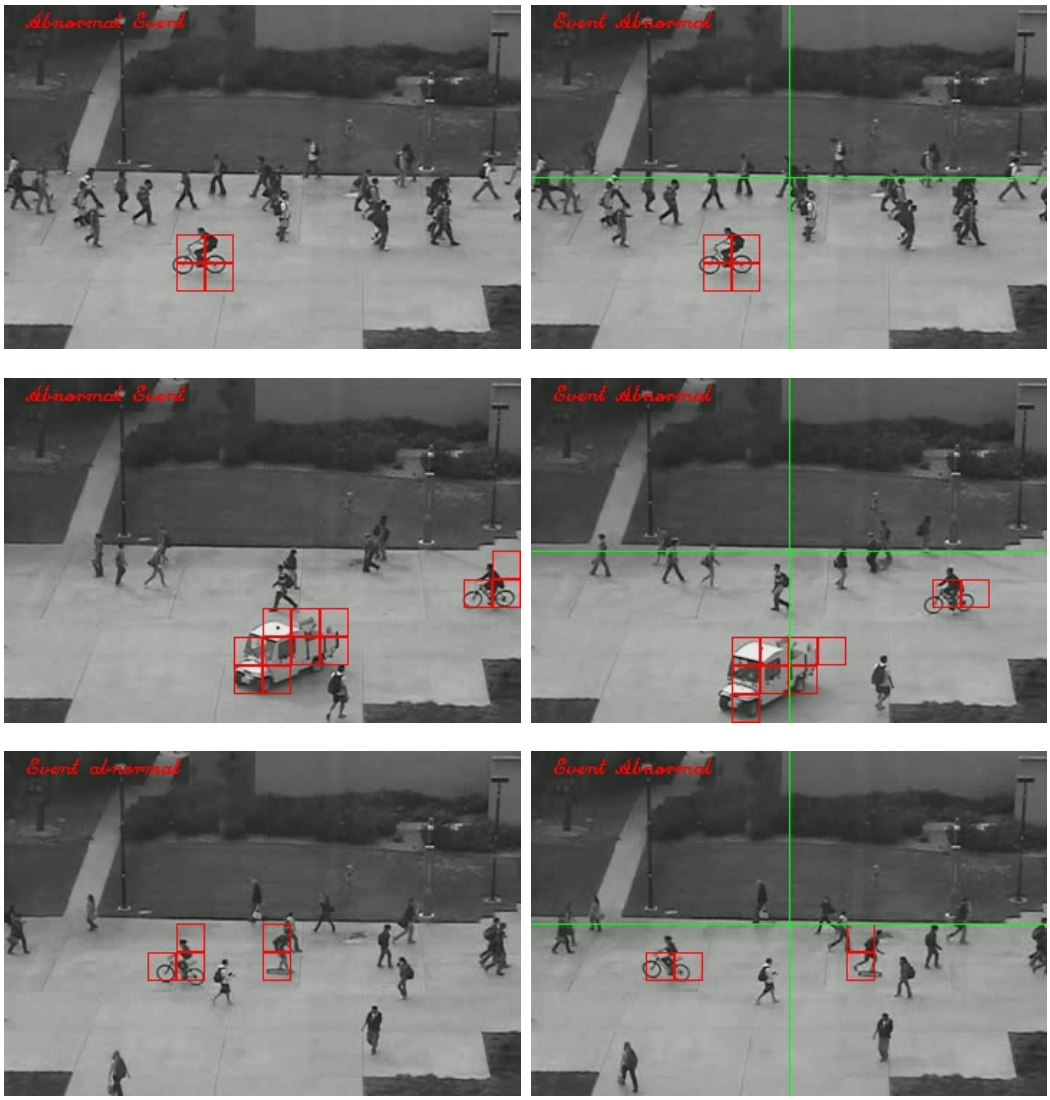


Figura 5.7: Resultados del método propuesto sobre la base de datos UCSD Peds2, en cada fila de imágenes se muestra la detección de los diferentes objetos como: carros, bicicletas y patinadores. La primera y la segunda columna de imágenes representa los resultados utilizando los dos tipos de clasificación propuestas, 01 y 04 regiones, respectivamente.

presentan problemas de perspectiva, tal es el caso de los vídeos de la base de datos UCSD Peds1. Sin embargo, para los vídeos que no presentan problemas de perspectiva la clasificación utilizando una sola región tiende a obtener mejores resultados que la clasificación utilizando cuatro regiones, tal es el caso de la base de datos UCSD Peds2.

En la Figura 5.8 se muestra las curvas ROC del método propuesto sobre la base de datos UCSD, utilizando los tipos de clasificación propuestos. Como se puede observar el rendimiento del método propuesto utilizando cuatro regiones de clasificación mejora a la clasificación utilizando una sola región en los videos que presentan problemas de

	01 región		04 regiones	
Métodos propuesto	EER	AUC	EER	AUC
UCSD Peds1	32.33 %	0.7386	**29.28 %**	**0.7923**
UCSD Peds2	**07.24 %**	**0.9778**	07.81 %	0.9771

Tabla 5.3: Resultados cuantitativos del método propuesto sobre la base de datos UCSD utilizando los dos tipos de clasificación propuestas (01 y 04 regiones). EL área bajo la curva (AUC) de la curva ROC y la tasa de igual error (EER) son calculados.

perspectiva.

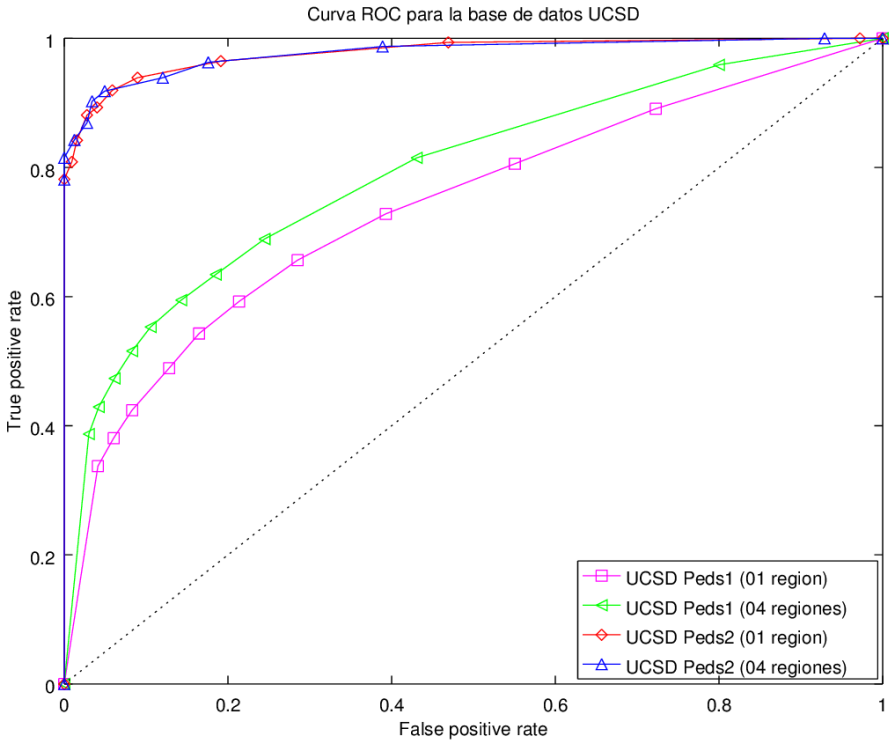


Figura 5.8: Las curvas ROC del método propuesto para la detección de EAL sobre la base de datos UCSD, utilizando los dos tipos de clasificación propuestas (01 y 04 regiones).

5.3.2.1. Comparación con otros Métodos de la Literatura

El rendimiento del método propuesto es comparado con varios métodos de la literatura tales como: el modelo de fuerza social (Mehran et al., 2009), el modelo de mezcla de texturas dinámicas (MDT) (Mahadevan et al., 2010), el método de Reddy (Reddy et al., 2011), las texturas de flujo óptico (Ryan et al., 2011), y el método HOFM

(Colque et al., 2015). En la Tabla 5.4 se muestra la comparación cuantitativa del método propuesto con los métodos mencionados. El rendimiento del método propuesto sobre la base de datos UCSD Peds1 logra un EER 29.2 % y un AUC de 0.792 que es comparable con algunos métodos de la literatura, pero algunos métodos nos superan tal es el caso de las texturas del flujo óptico (Ryan et al., 2011). El rendimiento del método propuesto sobre la base de datos UCSD Peds1 decae principalmente debido a que los vídeos de esta base de datos presentan problemas de perspectiva (Sección 4.3). Sin embargo, el rendimiento del método propuesto sobre la base de datos UCSD Peds2 logra un EER de 07.2 % y un AUC de 0.977 este resultado supera a todos los resultados de la literatura.

	Peds1		Peds2	
Métodos	EER	AUC	EER	AUC
Fuerza social (Mehran et al., 2009)	31.0 %	-	42.0 %	-
MDT (Mahadevan et al., 2010)	25.0 %	-	25.0 %	-
Reddy (Reddy et al., 2011)	22.5 %	-	20.0 %	-
Texturas (Ryan et al., 2011)	**23.1 %**	**0.838**	12.7 %	0.939
HOFM (Colque et al., 2015)	33.3 %	0.715	19.0 %	0.899
Método propuesto	29.2 %	0.792	**07.2 %**	**0.977**

Tabla 5.4: Comparación del rendimiento del método propuesto con los métodos de la literatura sobre la base de datos UCSD. EL área bajo la curva (AUC) de la curva ROC y la tasa de igual error (EER) son calculados.

5.4. Consideraciones Finales

EL rendimiento del método propuesto sobre las diferentes base de datos logra superar a los resultados de la literatura en los vídeos que no presentan problemas de perspectiva. Por otro lado, los resultados del método propuesto es comparable con algunos resultados de la literatura en los vídeos que presentan problemas de perspectiva. Para abordar el problema de perspectiva en esta tesis se propone la clasificación por regiones (01 y 04 regiones) logrando mejorar los resultados.

Capítulo 6

Conclusiones y Trabajos Futuros

En esta tesis se ha propuesto un modelo de detección de eventos anómalos basado en las características locales de movimiento y la apariencia. Las características de movimiento son adecuados para detectar eventos anómalos que tienen una velocidad de movimiento muy alta. Sin embargo, no todos los eventos anómalos poseen una velocidad alta. Para este tipo de eventos se introduce el uso de características de apariencia y así lograr detectar eventos anómalos con una velocidad normal. Las atributos utilizados en el modelo propuesto son extraídos de los parches espacio-temporales sin superposición de las secuencias de vídeo con el objetivo de detectar y localizar eventos anómalos locales tales como bicicletas, vehículos y patinadores. En la etapa de detección, se introduce el clasificador de la distancia mínima por regiones para determinar si una muestra de prueba es anómalo o no, utilizando un umbral predefinido. Además, cabe resaltar que el método propuesto utiliza un único modelo general para detectar ambos tipos de eventos anómalos globales y locales.

El método propuesto ha sido evaluado en varios conjuntos de bases de datos tales como UMN y UCSD, para verificar su rendimiento. Los experimentos muestran que los resultados del método propuesto son comparables con los trabajos de la literatura y en algunas bases de datos logra superarlas. Por ejemplo, en la base de datos UCSD Peds2 se logra un EER de 07.2 % y un AUC de 0.977 y en la base de datos UMN (escena 1 y 3) se obtiene un AUC de 0.998 y 0.995, respectivamente. Sin embargo los resultados decaen por debajo de los resultados de la literatura cuando los vídeos presentan problemas de perspectiva. Para abordar el problema de perspectiva en este tesis se propone la clasificación por regiones locales (01 y 04 regiones) logrando mejorar los resultados sobre la base de datos Peds1, de un EER de 32.3 % a 29.2 % y de un AUC de 0.738 a 0.792 y en la base de datos UMN (escena 2) de un AUC de 0.918 a 0.948, pero aún así todavía no se ha podido lograr superar los resultados de la literatura.

6.1. Limitaciones

Una de las principales limitaciones del método propuesto se presenta en la fase de extracción de características de los vídeos con problemas de perspectiva. Debido al problema de perspectiva algunos eventos anómalos no pueden ser detectados. Por ejemplo, cuando el evento anómalo esta más lejos de la cámara, la tasa de falsos negativos incrementa. Además, otra limitación en la etapa de detección de eventos anómalos es establecer manualmente el umbral para determinar si la muestra entrante es anómalo o no.

6.2. Trabajos futuros

Como trabajo futuro se enfocará extender la evaluación del método propuesto sobre otras bases de datos, experimentar con otras características y modelos de aprendizaje para detectar eventos anómalos en diferentes contextos. También se investigará acerca de los vídeos que presentan problemas de perspectiva para poder realizar de manera más eficaz la extracción de características y la clasificación de los eventos anómalos. Además, en la etapa de detección utilizar un umbral adaptativo para el clasificador de distancia mínima.

Bibliografía

Ali, S. and Shah, M. (2007). A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE.

Amin, A., Anzum, M. F., Mondol, M. H., et al. (2014). *Abnormal behavior detection of human by video surveillance system*. PhD thesis, BRAC University.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

Bouguet, J.-Y. (2001). Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 5:1–10.

Chan, A. B. and Vasconcelos, N. (2008). Modeling, clustering, and segmenting video with mixtures of dynamic textures. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(5):909–926.

Chaudhry, R., Ravichandran, A., Hager, G., and Vidal, R. (2009). Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1932–1939. IEEE.

Chen, D.-Y. and Huang, P.-C. (2013). Visual-based human crowds behavior analysis based on graph modeling and matching. *Sensors Journal, IEEE*, 13(6):2129–2138.

Colque, M., Hugo, R. V., Caetano, C. A., and Schwartz, W. R. (2015). Histograms of optical flow orientation and magnitude to detect anomalous events in videos. In *Graphics, Patterns and Images (SIBGRAPI), 2015 28th SIBGRAPI Conference on*, pages 126–133. IEEE.

Cong, Y., Yuan, J., and Liu, J. (2013). Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition*, 46:1851–1864.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Gu, P., Temizel, A., Temizel, T. T., et al. (2013). Real-time global anomaly detection for crowd video surveillance using sift. In *IET Conference Proceedings*. The Institution of Engineering & Technology.

Horn, B. K. and Schunck, B. G. (1981). Determining optical flow. In *1981 Technical symposium east*, pages 319–331. International Society for Optics and Photonics.

INEI (Junio de 2016). Instituto nacional de estadística e informática. https://www.inei.gob.pe/media/MenuRecursivo/boletines/informe-tecnico-n03_seguridad-ciudadana-ene-jun2016.pdf.

Kratz, L. and Nishino, K. (2009). Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1446–1453. IEEE.

Li, T., Chang, H., Wang, M., Ni, B., Hong, R., and Yan, S. (2015). Crowded scene analysis: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 25(3):367–386.

Li, W., Mahadevan, V., and Vasconcelos, N. (2014). Anomaly detection and localization in crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):18–32.

Mahadevan, V., Li, W., Bhalodia, V., and Vasconcelos, N. (2010). Anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1975–1981. IEEE.

Mehran, R., Oyama, A., and Shah, M. (2009). Abnormal crowd behavior detection using social force model. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 935–942. IEEE.

Nallaivarothayan, H., Fookes, C., Denman, S., and Sridharan, S. (2014). An mrf based abnormal event detection approach using motion and appearance features. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 343–348. IEEE.

Raghavendra, R., Bue, A. D., Cristani, M., and Murino, V. (2011a). Optimizing interaction force for global anomaly detection in crowded scenes. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 136–143. IEEE.

Raghavendra, R., Del Bue, A., Cristani, M., and Murino, V. (2011b). Abnormal crowd behavior detection by social force optimization. In *Human Behavior Understanding*, pages 134–145. Springer.

Reddy, V., Sanderson, C., and Lovell, B. C. (2011). Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In *CVPR 2011 WORKSHOPS*, pages 55–61. IEEE.

Roshtkhari, M. J. and Levine, M. D. (2013). An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer vision and image understanding*, 117(10):1436–1452.

Ryan, D., Denman, S., Fookes, C., and Sridharan, S. (2011). Textures of optical flow for real-time anomaly detection in crowds. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 230–235. IEEE.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Shi, Y., Gao, Y., and Wang, R. (2010). Real-time abnormal event detection in complicated scenes. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, pages 3653–3656. IEEE Computer Society.

Sodemann, A., Ross, M. P., Borghetti, B. J., et al. (2012). A review of anomaly detection in automated surveillance. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(6):1257–1272.

UCSD, a. d. d. (2013). Ucsd anomaly detection dataset. http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm.

UMN, c. d. (2012). Umn crowd dataset: Detection of unusual crowd activity. http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi.

Vapnik, V. and Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24.

Wang, B., Ye, M., Li, X., and Zhao, F. (2011). Abnormal crowd behavior detection using size-adapted spatio-temporal features. *International Journal of Control, Automation and Systems*, 9(5):905–912.

Wang, B., Ye, M., Li, X., Zhao, F., and Ding, J. (2012). Abnormal crowd behavior detection using high-frequency and spatio-temporal features. *Machine Vision and Applications*, 23(3):501–511.

Wang, T. and Snoussi, H. (2014). Detection of abnormal visual events via global optical flow orientation histogram. *Information Forensics and Security, IEEE Transactions on*, 9(6):988–998.

Wu, S., Moore, B. E., and Shah, M. (2010). Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2054–2060. IEEE.

Wu, S., Wong, H.-S., and Yu, Z. (2014). A bayesian model for crowd escape behavior detection. *Circuits and Systems for Video Technology, IEEE Transactions on*, 24(1):85–98.

Xiong, G., Wu, X., Chen, Y.-l., and Ou, Y. (2011). Abnormal crowd behavior detection based on the energy model. In *Information and Automation (ICIA), 2011 IEEE International Conference on*, pages 495–500. IEEE.

Xu, J., Denman, S., Sridharan, S., Fookes, C., and Rana, R. (2011). Dynamic texture reconstruction from sparse codes for unusual event detection in crowded scenes. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 25–30. ACM.

Yang, H., Cao, Y., Wu, S., Lin, W., Zheng, S., and Yu, Z. (2012). Abnormal crowd behavior detection based on local pressure model. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE.